

# Anwendung allgemeiner Klassifikationsverfahren zur Evaluation von in vitro-Systemen

H. Hecker und P. Wübbelt

## 1 Einführung

Die Evaluation von in vitro-Systemen ist in dem hier vorliegenden Kontext allgemein als der Versuch anzusehen, die Gesamtheit der bei den in vitro-Systemen gemessenen Daten einer Substanz zu den gleichzeitig vorliegenden Ergebnissen der R41-Klassifikation in Beziehung zu setzen (vgl. [6]). Aufgrund der Vielzahl der Substanzen, zu denen in diesem Sinne verwertbare Meßergebnisse vorliegen, wird dabei angestrebt, sowohl die Struktur dieser Beziehungen zu untersuchen als auch quantitative Aussagen hierüber zu ermöglichen. Letztlich wird die Etablierung einer möglichst guten *Klassifikationsregel* angestrebt: Auf der Basis der vorliegenden Daten (der "Lern"stichprobe) soll eine Zuordnungsregel erstellt werden, welche zukünftige Beobachtungen mit möglichst kleiner Fehlerrate derjenigen Klasse zuordnet, welche ein in vivo-Test derselben Substanz zuordnen würde. Neben der Etablierung einer solchen Regel wird aber auch eine Bewertung ihrer Klassifikationseigenschaften erforderlich, um darüber entscheiden zu können, ob sie den gestellten Anforderungen genügt.

Die hier allgemein formulierten Ziele und Kriterien können in sehr verschiedener Weise konkretisiert werden. Dieses geht bereits aus den Anwendungen der linearen und der nicht-parametrischen Diskriminanzanalyse (siehe [6]) hervor. Weitere, hiervon unterschiedliche Konkretisierungen wurden in zwei zusätzlich angewendeten multivariaten statistischen Verfahren zur Untersuchung der Beziehungen zwischen in vitro- und in vivo-Ergebnissen verfolgt. Der Grund für ihre Anwendung liegt in folgendem: Einerseits ist in Betracht zu ziehen, daß die Beziehungen zwischen beiden Variablenbereichen zur in vivo- und in vitro-Untersuchung in starkem Ausmaße *nicht-linear* und in hohem Grade von *Wechselwirkungen* geprägt sein könnten. Zum anderen ist zu berücksichtigen, daß Fehlklassifikationen verschiedener Art (Nicht-Erkennen der Toxizität oder fälschliche Einteilung als toxisch) im vorliegenden Kontext unterschiedlich zu bewerten sind. Dieses sollte sich auch in der Wahl der statistischen Analyse in geeigneter Weise widerspiegeln.

Weiterhin könnte die Zielsetzung einer Analyse auch dahingehend verändert werden, daß man nicht versucht, *jede* Substanz entweder als "toxisch" oder als "nicht-toxisch" zu klassifizieren, sondern daß man stattdessen nach *zwei* Clustern von Substanzen sucht, deren Zuordnung in diesem Sinne als sehr sicher anzusehen ist, und daß man darüber hinaus einen *dritten* Cluster von Substanzen identifiziert, bei denen eine solche Zuordnung nicht möglich erscheint.

Diese Aspekte werden –mit unterschiedlicher Gewichtung– in den Verfahren *CART* (siehe [2]) und *CBR* berücksichtigt. Beide Ansätze werden im Folgenden in jeweils einem eigenen Abschnitt vorgestellt und sodann auf die Auswertung der in vitro-Systeme angewendet.

## 2 Diskriminanzanalysen mit Hilfe von *CART*

### 2.1 Klassifikationsbäume: Grundzüge und spezielle Optionen von *CART* (Classification And Regression Trees)

In der Diskriminanzanalyse geht man von zwei verschiedenen Datensätzen aus, welche die einzelne Beobachtungseinheit kennzeichnen: Einerseits beschreibt die (kategorielle) Variable  $Y$  mit den möglichen Ausprägungen  $y = 0, 1, 2, \dots, K$ , die *Gruppe* oder *Klasse*, zu welcher diese Beobachtungseinheit gehört. Andererseits sind  $X_1, X_2, \dots, X_p$  weitere Variablen zur Kennzeichnung der Beobachtungseinheit. In der *Lernstichprobe* sind für jede Beobachtungseinheit (eventuell bis auf Missings) alle Variablen, auf jeden Fall jedoch die Variable  $Y$  und mindestens eine der Variablen des Vektors  $\mathbf{X} = (X_1, X_2, \dots, X_p)$  bekannt. In dieser Lernstichprobe besteht also prinzipiell die Möglichkeit, die (bedingte) Verteilung von  $\mathbf{X}$  in Abhängigkeit von der Klassenzugehörigkeit  $Y$  empirisch zu untersuchen:

$$\mathcal{L}(\mathbf{X}|Y = y) \quad (y = 0, 1, \dots, K) \quad (1)$$

Mit Hilfe der Kenntnis der "a priori-Verteilung" von  $Y$ :  $p_i = P(Y = i)$  ( $i = 0, 1, \dots, K$ ) wird daraus umgekehrt auf die (bedingte) Verteilung von  $Y$  in Abhängigkeit von der Wertekonstellation der Variablen  $X_1, X_2, \dots, X_p$  (die "a posteriori-Verteilung" von  $Y$  zu gegebenem  $\mathbf{X}$ ) geschlossen:

$$p_i(\mathbf{x}) = P(Y = i|\mathbf{X} = \mathbf{x} = (x_1, x_2, \dots, x_p)) \quad (2)$$

Dies wiederum ist die Basis für eine Regel zur Klassifikation von (zukünftigen) Beobachtungseinheiten, bei denen die Gruppenzugehörigkeit  $y$  unbekannt ist:

Zu den gegebenen Werten  $\mathbf{x} = (x_1, x_2, \dots, x_p)$  einer Beobachtungseinheit wird diese derjenigen Klasse  $y$  zugeordnet, bei welcher die berechnete a posteriori-Wahrscheinlichkeit maximal ist. (Dadurch wird gleichzeitig die *Wahrscheinlichkeit für Fehlklassifikationen* in der Gesamtpopulation minimiert).

Im Fall von nur  $K = 2$  Gruppen gelten, falls die Vektoren  $\mathbf{X}$  bedingt (d.h. gegeben  $Y = y$ ) normalverteilt sind mit identischer Kovarianzmatrix, für die a posteriori-Wahrscheinlichkeiten  $p_i(\mathbf{x})$  die Gleichungen (vgl.z.B. [1], S. 134)

$$\log \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (3)$$

Die einzelnen Variablen  $X_i$  gehen daher unter diesen Voraussetzungen *linear*, nämlich mit den Koeffizienten  $\beta_i$ , in die Berechnung der a posteriori-Wahrscheinlichkeiten und damit auch in die Klassifikationsregel ein. Darüber hinaus sind in diesem Modell *keine Wechselwirkungen* vorgesehen: Der Einfluß der Veränderung z.B. der Variablen  $X_1$  um eine Einheit beeinflusst das Verhältnis der a posteriori-Wahrscheinlichkeiten (in der logarithmierten Skala) *unabhängig von dem Stand der Variablen  $X_2, \dots, X_p$* .

Es kann in konkreten Anwendungen Gründe geben, von beiden Annahmen abzugehen:

1. Die *Linearität* wird nicht als gegeben angesehen (und es gibt andererseits keine begründeten Annahmen über spezielle Transformationen, welche eine Linearität erzeugen). Dies ist insbesondere immer dann der Fall, wenn nicht alle Variablen  $X_i$  als quantitativ sondern teilweise als ordinal skaliert oder qualitativ anzusehen sind.
2. Es wird als möglich angesehen, daß der "Einfluß" einer Variablen, z.B.  $X_1$  auf die a posteriori-Wahrscheinlichkeit  $p_i(x_1, x_2, \dots, x_p)$ , in Ausmaß (und Richtung) davon abhängen kann, wie die Konstellation der Werte der anderen Beobachtungsvariablen  $x_2, x_3, \dots, x_p$  ist.

Insbesondere zur Entdeckung solcher "Wechselwirkungen", dabei allerdings im Rahmen der multiplen Regression, wurde 1961 von MORGAN und SONQUIST das Baumanalyseprogramm *AID = Automatic Interaction Detection* entwickelt ([7]). In dieser Version ist  $Y$  also eine quantitative Variable, und das Programm sucht in einem schrittweisen Verfahren nach Untergruppen, innerhalb derer jeweils eine möglichst kleine Varianz bezüglich  $Y$  vorliegt und die sich andererseits bezüglich des Mittelwertes von  $Y$  möglichst stark voneinander unterscheiden. Diese Untergruppen müssen durch die Werte der  $X$ -Variablen charakterisierbar sein. Genauer geschieht dies dadurch, daß zunächst die gesamte Stichprobe und danach die weiter entstandenen Untergruppen durch jeweils eine Bedingung der Form " $X_i \leq x$ " (bei ordinalen Variablen  $X_i$ ) oder " $X_j \in \{c_{j1}, c_{j2}, \dots, c_{jJ}\}$ " (bei kategoriellen Variablen mit  $< J$  Kategorien  $c_j$ ) in *zwei* weitere Untergruppen geteilt werden, in denen diese Bedingung also erfüllt bzw. nicht erfüllt ist. Zu jedem solchen *Split* werden *alle Variablen  $X_i$*  und *alle in der angegebenen Form möglichen Bedingungen* überprüft und daraus diejenige ausgewählt, die nach den angegebenen Kriterien die beste weitere Aufteilung liefert. Eine solche rekursive Aufteilung wird üblicherweise als *binärer Baum* dargestellt, in denen die einzelnen Untergruppen als *Knoten* erscheinen und solche Knoten, die nicht weiter geteilt werden, speziell als *Endknoten* oder *Blätter*. Eine etwaige Wechselwirkung –wie oben beschrieben– zwischen  $X_1$  und  $X_2$  ist aus einer solchen Darstellung beispielsweise daran zu erkennen, daß der Knoten " $X_1 \leq x$ " durch die Variable  $X_2$  weiter gesplittet wird, während der Knoten " $X_1 > x$ " gar nicht oder durch eine andere Variable weiter geteilt wird. (Im ersten Knoten hat  $X_2$  dann einen starken und insgesamt den größten "Einfluß" auf  $Y$ , im zweiten jedoch keinen bedeutsamen oder zumindest –verglichen mit den anderen  $X$ -Variablen– nicht den größten Einfluß).

Die in diesem Ansatz entwickelten Grundideen wurden von BREIMAN, FRIEDMAN, OLSHEN und STONE wieder aufgenommen, in wesentlichen Punkten weiterentwickelt und 1984 unter dem Namen *CART = Classification And Regression Trees* sowohl als Monographie als auch als Computerprogramm veröffentlicht. Die Weiterentwicklung und Konsolidierung durch Breimann et al. dieses ursprünglich mehr heuristisch entwickelten Ansatzes betreffen u.a.:

1. Erweiterung auf Probleme der Diskriminanzanalyse (*Klassifikations-bäume*)
2. Einführung von Optimalitätskriterien für einzelne Splits und für den ganzen Baum
3. Lösung des Problems der Überklassifikation durch "Pruning" (Stutzen zu weit verzweigter Bäume nach vorgegebenen Kriterien)
4. Einbau der Kreuzvalidierung zur Erzeugung "verlässlicher" Bäume ("honest trees")

Zusätzliche, für Anwendungen wichtige Ergänzungen betreffen z.B:

1. Die Auswahl von "Surrogat"-Variablen für den Fall fehlender Werte bei einer Split-Variablen (dadurch können grundsätzlich *alle* Beobachtungseinheiten in die Analyse einbezogen werden).
2. Die Möglichkeit, die "*Kostenfunktion*" für die Belegung der verschiedenen Möglichkeiten zur Fehlklassifikation (im 2-Gruppen-Fall die falsch-positive und falsch-negative Klassifikation) flexibel zu steuern.

## 2.2 Anwendung von CART auf die R41- Klassifikation

Die Auswertung beruht auf den *in vitro*- und *in vivo*-Untersuchungen von  $N = 134$  Substanzen. Die detaillierte Darstellung der Messungen sowie die Beschreibung der für die Klassifikation benutzten Merkmale ist dem Artikel von S.GLASER im vorliegenden Tagungsband ([6]) zu entnehmen. Insbesondere wurden dieselben Variablen wie dort zur Beschreibung der *in vitro*-Messungen (die

” $X$ -Variablen) benutzt und ebenso dieselbe Zuordnung der Substanzen in die Kategorie ”  $R_{41}$ ” mit der Kurzbezeichnung ”*reizende*” ( $n = 48$ ) Substanz bzw. in die Kategorie ”*nicht reizende*” Substanz ( $n = 86$ ). Lediglich auf die logarithmierten Versionen der Meßgrößen wurde verzichtet, da von *CART* ohnehin nur das Ordinalniveau der Skalen ausgenutzt wird.

### Kreuzvalidierung

Die Baumanalysen wurden mit 10-facher Kreuzvalidierung durchgeführt. Dies bedeutet, daß zur Ermittlung der Fehlklassifikationsraten ein zufällig ausgewählter Anteil von 10% der Stichprobe vor der Erstellung des Klassifikationsbaumes eliminiert wurde, und daß die auf den verbleibenden 90% der Fälle basierende Klassifikationsregel auf diese zuvor eliminierten Fälle angewendet wurde. Insgesamt wurde diese Prozedur 10 mal durchgeführt, wobei die Auswahl der jeweils eliminierten Fälle variierte, so daß am Ende die gesamte Stichprobe betroffen war.

Bei diesem Vorgehen wird die Fehlklassifikationswahrscheinlichkeit ohne systematische Verzerrung geschätzt (und dadurch im Schnitt *nicht unterschätzt*, wie es bei reiner ”Resubstitution”, d.h. bei Anwendung der Klassifikationsregel auf genau die Fälle, die zur Erstellung der Klassifikationsregel benutzt wurden, zu erwarten ist).

### ”Kosten” für Fehlklassifikation

Anstelle der Fehlklassifikations*wahrscheinlichkeit* ist es häufig auch sinnvoll, die Fehlklassifikations*kosten* zur Bewertung einer Klassifikationsregel zu benutzen. Dadurch erhält man die Möglichkeit, verschiedene Arten von Fehlklassifikationen unterschiedlich stark zu ”bestrafen”. Im Falle der Entscheidung über die Zugehörigkeit einer Substanz zur Klasse  $R_{41}$  läßt sich dieses in der folgenden Tabelle einfach darstellen:

Tatsächliche Klasse:	Klassifikation als:	”Kosten”:
0 = nicht reizend	1 = $R_{41}$ (”reizend”)	$K_{01}$
1 = $R_{41}$ (”reizend”)	0 = nicht reizend	$K_{10}$

Setzt man im Fall einer *richtigen* Klassifikation noch die Kosten auf den Wert *Null*, so errechnen sich daraus die zu erwartenden Kosten einer Klassifikationsregel  $R$  insgesamt als

$$E(K) = K_{01} P(R = 1|Y = 0) P(Y = 0) + K_{10} P(R = 0|Y = 1) P(Y = 1) \quad (4)$$

wobei mit  $P(R = i|Y = j)$  die bedingte Wahrscheinlichkeit für die Zuordnung zur Klasse  $i$  bezeichnet ist, falls die Substanz tatsächlich zur Klasse  $j$  gehört ( $i, j = 0, 1$ ).

### ROC-Kurven

In den folgenden Auswertungen wurde das Verhältnis der Fehlklassifikationskosten:

$$\rho = \frac{K_{10}}{K_{01}} \quad (5)$$

variiert. Der Grund hierfür liegt darin, daß im vorliegenden Kontext das fälschliche Übersehen einer toxischen Substanz als schwerwiegender angesehen wird als der umgekehrte Fehler; dieses verlangt einen höheren Wert für  $K_{10}$  als für  $K_{01}$ :  $\rho$  ist  $> 1$  zu wählen. Darüber hinaus erhält

man in Abhängigkeit von der Wahl von  $\rho$  unterschiedliche (bedingte) Fehlklassifikationswahrscheinlichkeiten und damit auch bedingte Wahrscheinlichkeiten für richtige Klassifikation:

$$P(R = 1|Y = 1) = \text{Sensitivität} \quad (6)$$

$$P(R = 0|Y = 0) = \text{Spezifität} \quad (7)$$

Eine *systematische* Variation von  $\rho$  ermöglicht es daher, das Trennvermögen des Datensatzes *insgesamt* in Form einer ROC-Kurve (welche zu jedem  $\rho$  das Wertepaar von Sensitivität und Spezifität als einen Punkt im  $(0,1) \times (0,1)$ -Koordinatenkreuz enthält) darzustellen. Eine kontinuierliche Veränderung von  $\rho$  erzeugt dabei jedoch nur zu einzelnen, diskreten Werten jeweils einen neuen Klassifikationsbaum, so daß die ROC-„Kurve“ hier nur aus einzelnen Punkten besteht.

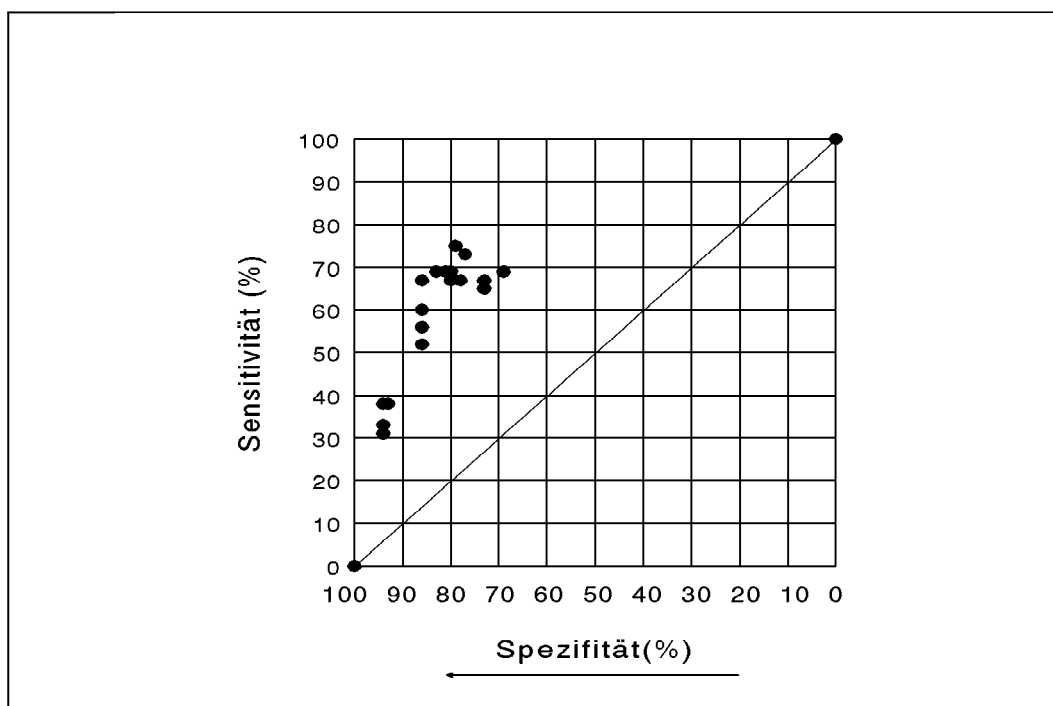


Abb. 1: ROC-Kurve für die *CART*-Analyse bei systematischer Variation der Kostenfunktion

Das Ergebnis dieser Auswertung ist in Abb. 1 dargestellt. Es kann wie folgt interpretiert werden:

1. Durch *CART* wird eine größere Anzahl von Klassifikationsbäumen generiert, die auch in der unverzerrten Bewertung durch die Kreuzvalidierung ein deutlich positives Trennvermögen aufweisen (Vergleich mit der Diagonalen, die durch reine Zufallszuordnung, d.h. ohne Kenntnis der *X*-Variablen erreichbar ist).
2. Zieht man nur die nicht-trivialen Klassifikationsbäume (in denen also nicht die Sensitivität oder die Spezifität gleich Null ist) in Betracht, so liegt die höchste erreichbare Spezifität bei 94%, die höchste erreichbare Sensitivität jedoch nur bei 75%. Werden Sensitivität und Spezifität gleich bewertet, so ist der optimale Baum (Sensitivität + Spezifität = maximum) durch den Punkt mit den Werten 75% und 79% für Sensitivität und Spezifität gekennzeichnet.
3. Bei Auswahl dieses Baumes und Anwendung der entsprechenden Klassifikationsregel müßte also damit gerechnet werden, daß 25% der R41-Substanzen nicht als solche erkannt und daß gleichzeitig 21% der nicht-reizenden Substanzen als R41-Substanz klassifiziert würden.

4. Die Ergebnisse für Sensitivität und Spezifität sind größenordnungsmäßig mit denen der linearen und der nicht-parametrischen Diskriminanzanalyse (siehe [6]) vergleichbar.

### ”Pruning”

Für die weitere Darstellung der Ergebnisse wurde der oben näher charakterisierte ”optimale” Baum mit 75% Sensitivität und 79% Spezifität (jeweils ermittelt durch Kreuzvalidierung) ausgewählt. Die oben beschriebene Prozedur der Kreuzvalidierung wird bei *CART* gleichzeitig dazu benutzt, um zu entscheiden, welche Verzweigungen des Baumes im nachhinein wieder rückgängig gemacht werden müssen (”Stutzen” oder ”pruning” des Baumes), da sie eine Überanpassung an die Daten der Stichprobe darstellen und sich bei der Zuordnung neuer Fälle als Verschlechterung der Klassifikationsregel erweisen.

Diese Regel über das Stutzen zu weit verzweigter Bäume ist nicht ganz eindeutig. Dies liegt daran, daß die Schätzung der Fehlklassifikationskosten (durch die Kreuzvalidierung) mit zufallsbedingten Unsicherheiten verbunden ist. Bei Berücksichtigung dieser Streuungen kann man –unter Hinnahme geringer Abweichung von den minimalen (geschätzten) Fehlklassifikationskosten– deutlich einfachere Klassifikationsbäume erhalten. Die drei nachfolgend dargestellten Klassifikationsbäume (Abb. 2, 5 und 6) entsprechen Fehlklassifikationskosten von 58%, 65% und 69% im Vergleich zur ”Zuordnung ohne Information”, bei welcher alle Substanzen (aufgrund ihres erhöhten Anteils in der Stichprobe; s.o.) in die Klasse ”nicht reizend” eingestuft werden. Vom optimalen Baum (53% relative Fehlklassifikationskosten) sind sie 1.0, 1.5 bzw. 2.0 Standardabweichungen ( $s = 8.7\%$ ) der relativen Kosten entfernt.

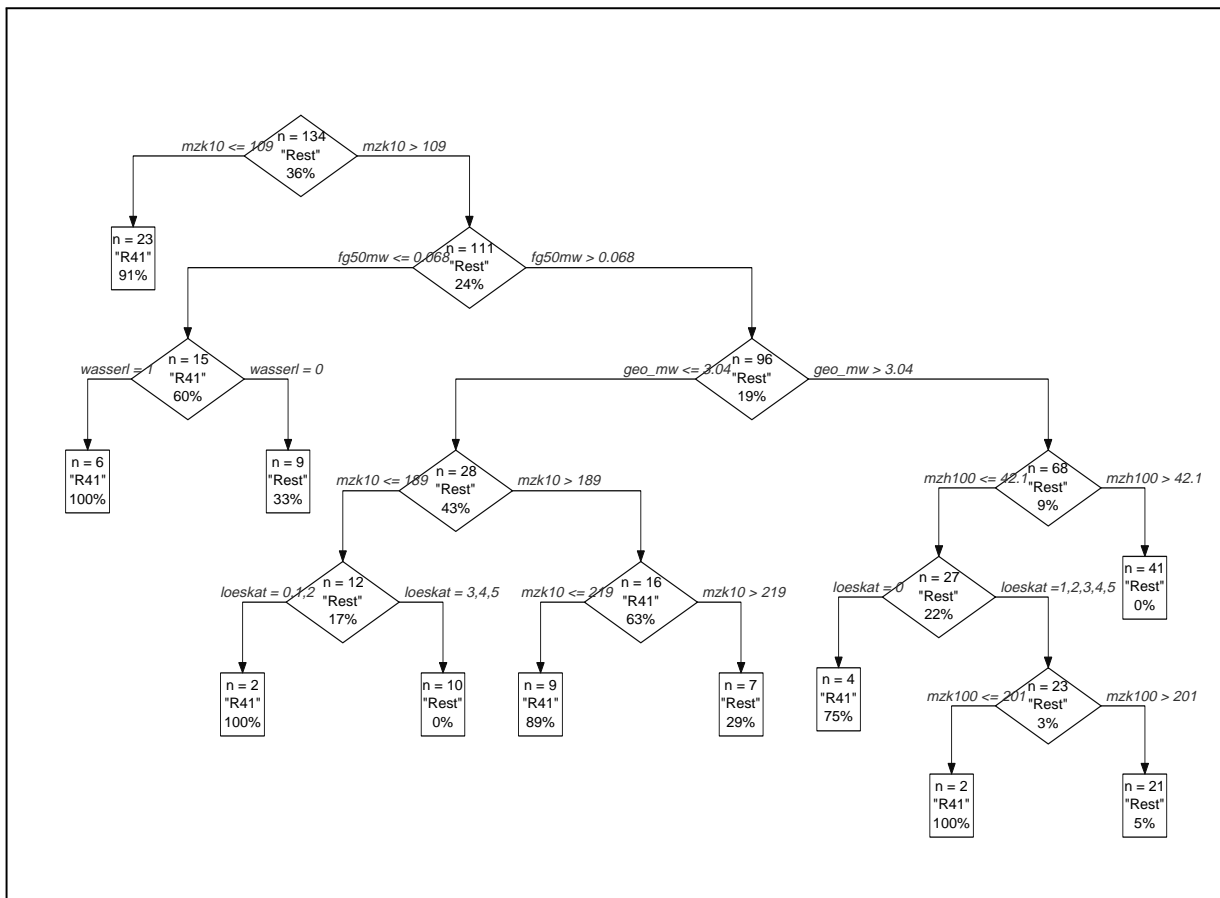


Abb. 2: Klassifikationsbaum mit relativen Fehlklassifikationskosten von 58% (mit Fallzahl, Zuordnung und Anteil R41-Substanzen in den einzelnen Knoten)

## Interpretation

In jedem Knoten des Baumes ist zunächst die *Anzahl der Substanzen* angegeben, die darin enthalten sind, die also den Bedingungen genügen, durch welche der jeweilige Knoten charakterisierbar ist. Die zweite Zahl gibt jeweils den Prozentsatz derjenigen Substanzen in dem Knoten an, die im in vivo-Test als "R41", als "reizend" klassifiziert wurden. Durch die sukzessive Aufspaltung sollte die Zuordnung zu der Klasse "reizend" oder "nicht-reizend" möglichst eindeutig sein, d.h. die Prozentsätze sollten möglichst nahe bei 0% oder 100% liegen. (Dies wird in [2] genauer durch die "impurity"-function definiert.)

Durch die erste Aufspaltung wird die Gesamtstichprobe mit insgesamt 36% R41-Substanzen in zwei Untergruppen (Knoten) mit 91% bzw. 24% R41-Substanzen geteilt. Der erste Knoten wird dabei durch die Bedingung " $mzk10 < 109$ " charakterisiert. Er umfaßt also genau die Substanzen, deren "Mittelwert der Reaktionszeit bis zur Koagulation bei Reizindex 10%" ( $mzk10$ ) kleiner als 109 ist. Es handelt sich dabei um 23 Substanzen, von denen 21 im in vivo-Test als R41 klassifiziert wurden. Dieser Knoten wird nicht weiter gesplittet: er kann nicht durch einen weiteren Split in zwei wesentlich "reineren" Knoten aufgeteilt werden. In der Klassifikationsregel des Baumes werden alle Substanzen, die aufgrund ihrer in vitro-Testergebnisse in diesen Knoten fallen, als "reizend" klassifiziert. Dieses ist durch die entsprechende Kennzeichnung unter dem (rechteckig dargestellten) Endknoten bezeichnet.

Der zweite Knoten des ersten Splits kann im Gegensatz zum ersten Knoten durch einen weiteren Split in einen linken Knoten mit einer deutlich *höheren* (60%) und einen rechten mit einer *niedrigeren* R41-Rate (19%) aufgeteilt werden. Der linke Knoten bildet dabei wiederum einen Endknoten. Er ist insgesamt dadurch charakterisierbar, daß er alle Substanzen mit *hohem  $mzk10$ -Wert* ( $mzk10 \geq 109$ ) und gleichzeitig *niedrigem  $fg50mw$ -Wert* ( $fg50mw < 0.068$ ) umfaßt ( $fg50mw$  ist der Mittelwert der  $fg50$ -Werte pro Substanz, wobei die EC50-Werte für Neutralrot mit einem Programm FITGRAPH ermittelt wurden). Substanzen in diesem Endknoten würden in zukünftigen Untersuchungen als "R41" klassifiziert. In der vorliegenden Lernstichprobe enthält er 15 Substanzen, von denen im tatsächlich durchgeführten in vivo-Test 9 Substanzen als R41 eingestuft wurden.

Eine genauere Untersuchung der Variablen  $fg50mw$  zeigt, daß ihr Trennvermögen (bezüglich der Klasse R41) nicht nur deshalb erst nach dem ersten Split "zum Zuge kommt", weil in der Gesamtstichprobe die Variable  $mzk10$  diesbezüglich dominant ist. Vielmehr kann  $fg50mw$  im Knoten " $mzk10 \geq 109$ " tatsächlich sehr viel besser als in der Gesamtpopulation zur Trennung der R41-Substanzen beitragen. Dies geht aus den beiden Abbildungen 3 und 4 hervor, in denen die kumulative Verteilung dieser Variablen jeweils für beide Klassen einmal in der Gesamtstichprobe und dann im genannten Knoten dargestellt ist. Insgesamt ist hierin ein Beispiel für eine Wechselwirkung zu sehen, und zwar speziell der Variablen  $mzk10$  und  $fg50mw$  in bezug auf die Klassifikationsvariable  $Y$ .

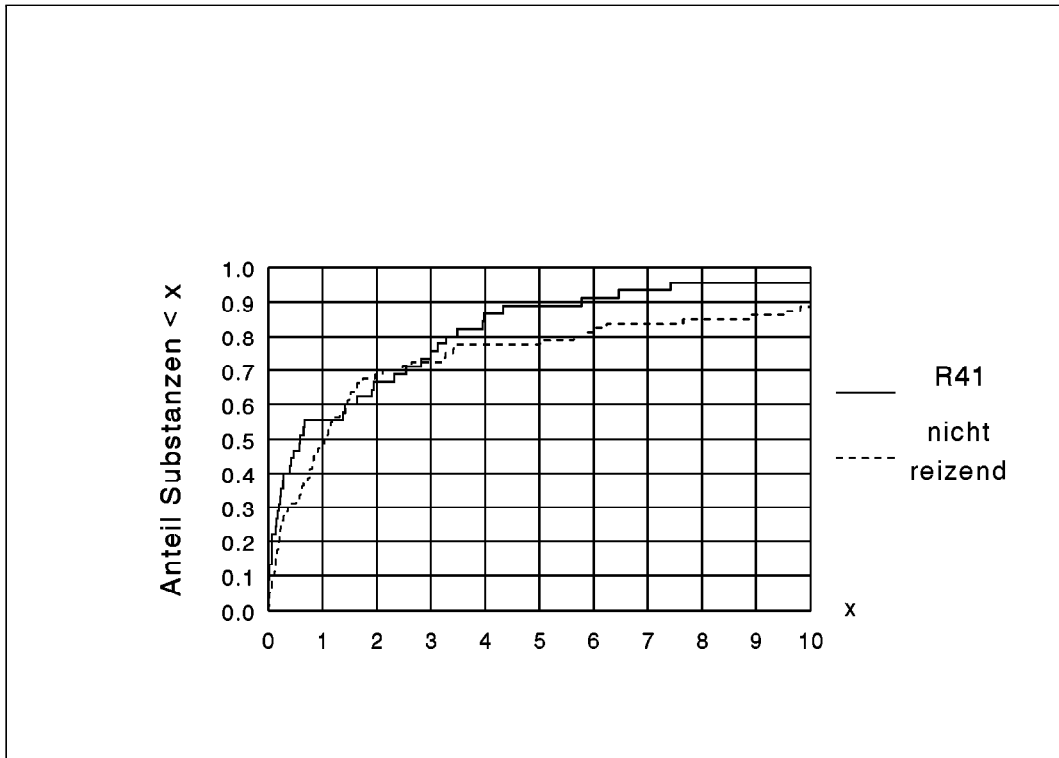


Abb. 3: Kumulative Verteilung von fg50mw in der Gesamtstichprobe

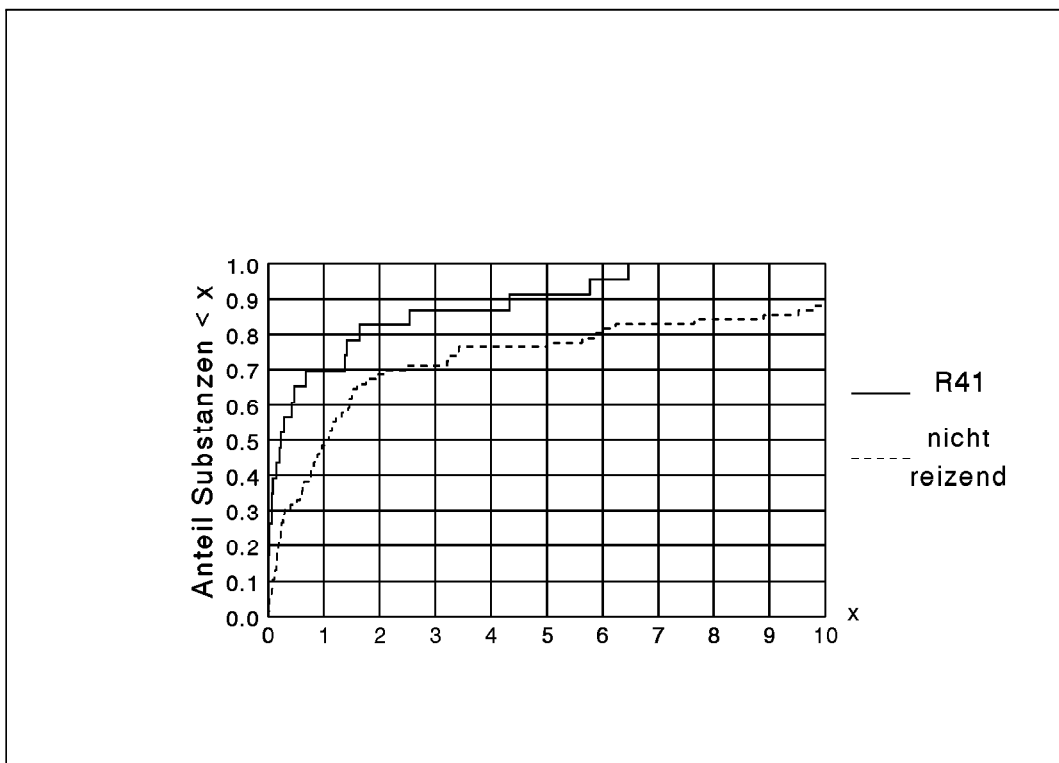


Abb. 4: Kumulative Verteilung von fg50mw für Substanzen mit erhöhtem mzk10-Wert

In der hier angegebenen Weise kann der Baum bis in alle seine Endknoten hin verfolgt und interpretiert werden. Insgesamt fällt dabei u.a. auf, daß die Endknoten teilweise sehr schwach besetzt sind. Wenn in einem solchen Fall zusätzlich die Charakterisierung des Endknotens durch

die Split-Variable einerseits und durch ihre R41-Häufigkeitsrate andererseits nicht mit den qualitativen Vorinformationen über die untersuchten Zusammenhänge übereinstimmt, so wird man solche Splits eher als "zufallsbedingt" einschätzen. Dies führt dazu, daß man nach den oben beschriebenen Überlegungen durch die Berücksichtigung der 1- bis 2-fachen Streuung der geschätzten Fehlklassifikationskosten zu weiterem "Stutzen" des hier beschriebenen Baumes kommt und damit zu den in Abbildungen 5 und 6 dargestellten vereinfachten Klassifikationsbäumen.

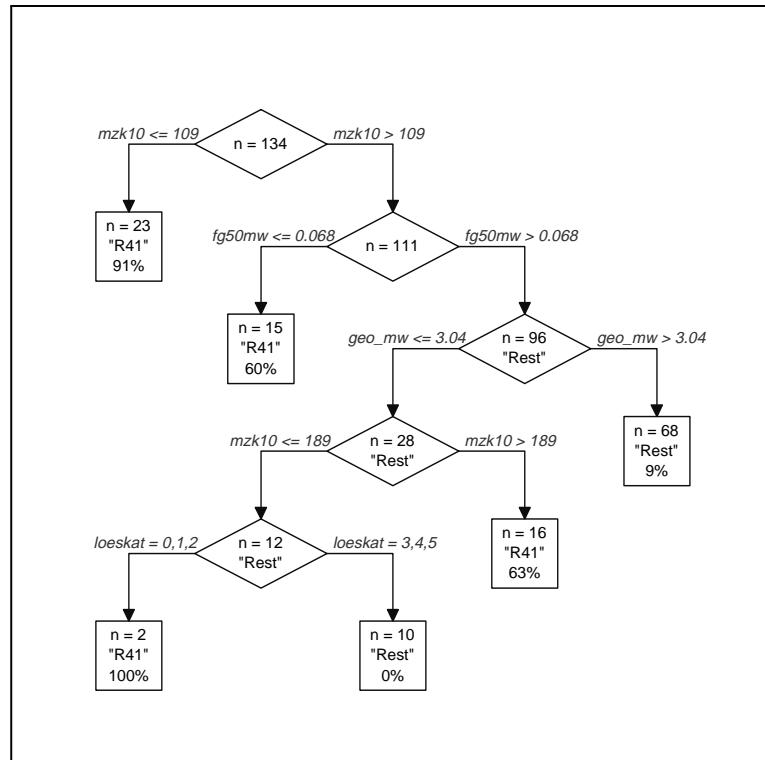


Abb. 5: Klassifikationsbaum mit relativen Fehlklassifikationskosten von 65% (mit Fallzahl, Zuordnung und Anteil R41-Substanzen in den einzelnen Knoten)

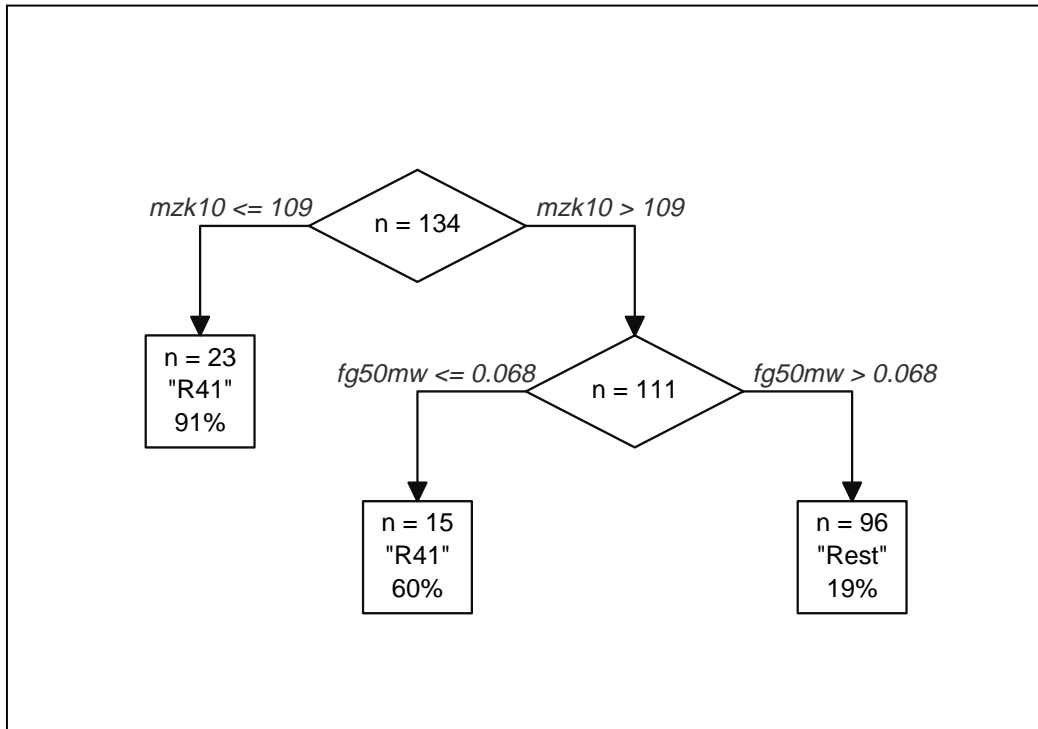


Abb. 6: Klassifikationsbaum mit relativen Fehlklassifikationskosten von 69% (mit Fallzahl, Zuordnung und Anteil R41-Substanzen in den einzelnen Knoten)

### Anwendung eines Klassifikationsbaumes

Die Anwendung eines Klassifikationsbaumes auf neue Substanzen hat gegenüber der linearen Diskriminanzanalyse den Vorteil, daß nicht grundsätzlich alle Variablen des Klassifikationsbaumes tatsächlich erhoben werden müssen (in der linearen Diskriminanzanalyse müssen alle Variablen der Diskriminanzfunktion zur Verfügung stehen). Ist beispielsweise die "Reaktionszeit bis zur Koagulation bei Reizindex 10%" ( $mzk_{10}$ ) kleiner als 109 sec., so steht nach den Entscheidungsbäumen der Abbildungen 2, 5 und 6 ihre Klassifizierung als "R41" bereits fest, so daß keine weiteren Messungen erforderlich sind.

Bei *fehlenden* Daten ist in der linearen Diskriminanzanalyse keine Zuordnung möglich; *CART* bietet für solche Fälle jedoch *Ersatzvariable* ("surrogate") an. Beispielsweise wird beim ersten Split des Baumes, wenn die Variable  $mzk_{10}$  nicht erhoben werden kann, die Bedingung " $mzk_{10} < 109$  sec." durch die Bedingung " $mzh_{10} < 24$  sec." bzw. durch " $mzh_{100} < 6.73$  sec." ersetzt ( $mzh_{10}$  und  $mzh_{100}$  bezeichnen den Mittelwert der Reaktionszeit bis zur Hämorrhagie bei Reizindex 10% bzw. 100%).

### Fehlklassifikationsraten

Abschließend hierzu ist noch zu bemerken, daß aus den Angaben der Abbildungen 2, 5 und 6 die Fehlklassifikationsraten *nach der Resubstitutionsmethode* berechnet werden können, die man also dadurch erhält, daß man den Klassifikationsbaum auf genau die Substanzen anwendet, die bereits zur Erstellung des Baumes benutzt wurden. Diese Raten erscheinen stets günstiger als die durch die oben beschriebene Methode der Kreuzvalidierung ermittelten Raten. Realistischer ("unverzerrt") sind jedoch die Ergebnisse der Kreuzvalidierung. In ihre Berechnung geht –neben der Suche nach den besten Schwellenwerten  $x$  für die Splits der Form  $X < x$  der einzelnen  $X$ -Variablen– sowohl die *Variablenselektion* ein, die mit der Auswahl eines Baumes verbunden ist, als auch die Ersetzung von fehlenden Werten durch *Surrogatvariable*. Bei zukünftiger Anwendung unter gleichen Bedingungen wie in der Lernstichprobe ist also mit Fehlklassifikationsraten zu rechnen, wie sie durch die Kreuzvalidierung geschätzt werden.

### 3 Analyse der Beziehungen zwischen in vitro– und in vivo–Daten mit Hilfe von *CBR*

#### 3.1 Clustern nach Response (*CBR*): Allgemeine Voraussetzungen und Zielsetzung des Verfahrens

Im Gegensatz zu diskriminanzanalytischen Verfahren, welche zunächst von den (bedingten) Verteilungen der  $X$ -Variablen  $\mathbf{X} = (X_1, X_2, \dots, X_p)$  in Abhängigkeit von der Klassenzugehörigkeit  $Y = 0, 1, \dots, K$  einer Beobachtungseinheit ausgehen und auf dieser Basis die a posteriori–Verteilung der Klassenvariablen  $Y$  in Abhängigkeit von  $\mathbf{X}$  berechnen, betrachtet *CBR* von vornherein  $\mathbf{X}$  als "unabhängige" oder "Prädiktorvariable" und  $Y$  als die "Zielvariable" oder den "Response". Bezogen auf den hiesigen Kontext lautet dabei die Fragestellung:

Gegeben die Ergebnisse der in vitro–Tests zu einer Substanz: Mit welcher Wahrscheinlichkeit wird das Ergebnis des in vivo–Tests positiv (R41) sein?

Aus dieser Fragestellung geht weiterhin hervor, daß dabei – wiederum im Gegensatz zur Diskriminanzanalyse – nicht die *Zuordnung* einer Substanz zu einer der vorgegebenen Klassen  $Y = 0, 1, \dots, K$  im Vordergrund steht. Vielmehr geht es darum, überhaupt nach *Unterschieden in der Wahrscheinlichkeitsverteilung von  $Y$*  zu suchen, die durch Unterschiede in den Werten der Prädiktoren  $X_1, X_2, \dots, X_p$  zu erklären sind.

Für *CBR* ist daher beispielsweise auch eine Unterscheidung von Untergruppen  $C_1$  und  $C_2$  relevant, in denen die Wahrscheinlichkeiten für den Response  $Y = 1$  (also z.B. für Klasse R41) deutlich verschieden, also z.B. 75% und 90% sind, *auch wenn die Diskriminanzanalyse beide Gruppen in dieselbe Klasse einteilen würde*.

Etwas allgemeiner formuliert lautet die Fragestellung bei *CBR*:

Gibt es Cluster von Fällen, die jeweils durch eine gut beschreibbare *Klasse von Prädiktorwerten* gekennzeichnet sind und für die jeweils gilt:

- *Innerhalb* eines Clusters ist die Verteilung der Response–Variablen *homogen*, und
- *zwischen verschiedenen Clustern* sind die Verteilungen der Response–Variablen sehr *unterschiedlich*.

Eine grundsätzliche Einschränkung für *CBR* betrifft den Typ der Prädiktorvariablen: Es wird angenommen, daß alle Prädiktoren *binär* sind (also nur die Werte 0 oder 1 annehmen können). Sie werden in dieser Form häufig als *Indikatoren für das Vorhandensein eines Risikofaktors* interpretiert. Bei der Untersuchung quantitativer Prädiktoren muß daher zunächst eine solche Dichotomisierung vorgenommen werden, wobei man z.B. auf bekannte Normgrenzen oder auch auf die Trennpunkte zurückgreifen kann, die sich aus einer *CART*–Analyse ergeben.

Die Responsevariable, die im vorliegenden Anwendungsfall ebenfalls binär ist, kann allgemein eine beliebige kategorielle, ordinale, quantitative oder auch zensierte Variable sein (weitere Variablentypen sind in [3] und [5] beschrieben).

Von den gesuchten Clustern wird bei diesem Verfahren *nicht* angenommen, daß sie als die Endknoten eines Verzweigungsbaumes dargestellt werden können. Stattdessen geht man davon aus, daß zu jedem Cluster ein oder zwei bestimmte Wertekonstellationen der Prädiktorvariablen, sogenannte *Muster* gehören, mit denen alle Fälle dieses Clusters exakt oder mit einer bestimmten

”Trefferzahl” übereinstimmen. Über die maximale Komplexität dieser Muster kann der Benutzer zusätzliche Bedingungen einführen und dadurch – je nach Vorwissen und nach Größe der vorliegenden Stichprobe – nach komplexeren oder eher nach einfach strukturierten Clusterlösungen suchen.

Das Ergebnis einer *CBR*-Analyse ist eine *Partition* der Menge aller möglichen Wertekombinationen der Prädiktoren in zwei oder mehrere durch *Muster* beschreibbare ”*Konstellationen*” sowie eine Schätzung der Verteilung (oder deren Parameter) der Responsevariablen *Y* innerhalb jeder dieser *Konstellationen*. (Auch *CART* bildet in diesem Sinne eine ”*Partition*”, jedoch sind die resultierenden ”*Konstellationen*” dort von anderer Struktur.) Die Gesamtheit aller *Fälle*, die einer *Konstellation* entsprechen, bilden das zugehörige *Cluster*.

Die Bewertung einer *Partition* in Hinblick auf das Ziel, Cluster mit möglichst unterschiedlicher Verteilung der Response-Variablen zu identifizieren, erfolgt über den *P*-Wert eines Signifikanztests: *Kleinere P-Werte* zu einem ausgewählten Test über die Nullhypothese identischer Verteilungen von *Y* in den Clustern einer *Partition* werden als *größere Unterschiede* in den Verteilungen zwischen den *Konstellationen* interpretiert.

Im Gegensatz zu *CART* erfolgt die Suche nach einer *Partition* nicht rekursiv sondern *erschöpfend*: *CBR* sucht unter *allen zugelassenen* *Partitionen* nach derjenigen mit dem minimalen *P*-Wert. Die Lösung ist daher jeweils *global* optimal (bei *CART* wird nur *lokal* jede einzelne Verzweigung optimiert).

Weitere Eigenschaften und Optionen zum Verfahren und zum Computerprogramm *CBR* findet man in [5]. Einführung und Programmbeschreibung sowie das Computerprogramm selber stehen im Verzeichnis `pub/MHH/BIOMETRIE` des servers `ftp.UNI-Hildesheim.de` zur Verfügung.

## 3.2 Anwendung von CBR auf die R41-Klassifikation

Die Analyse der Daten zur in vivo- und in vitro- Testung der vorliegenden Substanzen mit Hilfe von *CBR* stützt sich auf Teilergebnisse der vorangegangenen Auswertung mit dem Programm *CART*: Da in *CBR* nur binäre Prädiktoren zugelassen sind, benutzten wir die durch *CART* vorgeschlagenen Trennpunkte zur Dichotomisierung der Prädiktoren. Weiterhin beschränkten wir uns auf die Berücksichtigung von 4 Prädiktoren, wobei sich deren Auswahl auf die in der *CART*-Analyse resultierenden Klassifikationsbäume stützt. Die für *CBR* benutzten binären Prädiktoren waren somit:

1.  $X_1$ :  $mzk10 \leq 109$ , d.h. ”Mittelwert der Reaktionszeit bis zur Koagulation bei Reizindex 10%” kleiner oder gleich 109 sec. ?
2.  $X_2$ :  $fg50mw \leq 0.068$ , d.h. ”Mittelwert der EC50-Werte für Neutralrot nach FITGRAPH” kleiner oder gleich 0.068 ?
3.  $X_3$ :  $geomw \leq 3.0?$ , d.h. ”geometrischer Mittelwert der Reizschwelle des HETCAM-Test” kleiner oder gleich 3.0?
4.  $X_4$ :  $mzh100 \leq 42.1$ , d.h. ” Mittelwert der Reaktionszeit bis zur Hämorrhagie bei Reizindex 100%” kleiner oder gleich 42.1 sec?

Alle vier Prädiktoren wurden mit ”0 = Nein” und ”1 = Ja” kodiert.

Zunächst werden hier also die von *CART* gefundenen, für die *einzelnen Verzweigungen optimalen* Splits zur Dichotomisierung der Prädiktoren übernommen. Es soll jedoch überprüft werden, ob sich damit Kombinationen von Prädiktorwerten bilden lassen, die in der *globalen* Sicht eine noch engere Beziehung zwischen in vitro- und in vivo-Daten aufzeigen als es durch die Beschränkung auf Baumstrukturen möglich ist. Insbesondere wird es zunächst darum gehen, 2-er-Partitionen zu finden, die zu einer Klasseneinteilung mit höherer Sensitivität führen als es bei *CART* der Fall war. Zum anderen wird gezielt nach einer Klassifikation gesucht, bei welcher die "sicher reizenden" und die "sicher nicht-reizenden" Stoffe identifiziert werden. Hierzu werden Partitionen in *drei* Cluster gebildet, wobei zwei Konstellationen der in vitro-Daten erwartet werden die entweder mit hoher Wahrscheinlichkeit der Klasse *R41* ("reizend") oder der Komplementärklasse "nicht reizend" zuzuordnen sind; gleichzeitig sollte die dritte Klasse mit fraglicher Zuordnung möglichst klein sein.

Als zusätzliche Einschränkung der Menge aller in *CBR* zugelassenen Konstellationen wird vorausgesetzt, daß die "Muster", durch welche die Klassen zu beschreiben sind, *homogen* sein sollen (s.u.). Damit wird berücksichtigt, daß bei allen Prädiktoren in gleicher Weise angenommen wird, daß die Klasse *R41* tendenziell eher mit dem Wert 1 des Prädiktors (niedrige Reaktionszeiten, Reizschwellen bzw. ED-50-Werte) assoziiert ist.

### 3.2.1 Partition in zwei Cluster

Die von *CBR* gefundenen 2-er Partitionen wurden in dieser Anwendung (anders als nach dem sonst üblichen *P*-Wert-Kriterium) danach sortiert, mit welcher *Sensitivität* in der Lernstichprobe sie verbunden sind, wenn man sie im Sinne der Diskriminanzanalyse als Zuordnungsregel benutzt. Diese Zuordnungsregel ist dabei gleichzusetzen mit der *Charakterisierung* der beiden Cluster der Partition *durch die Konstellationen der Prädiktorwerte*. Die optimale Partition sieht danach wie folgt aus:

Cluster- Nummer	$X_1$	$X_2$	$X_3$	$X_4$	$t$	$n$	<i>R41</i>		<i>Rest</i>	
							abs.	(%)	abs.	(%)
1	1	1	1	1	1	85	44	(52%)	41	(48%)

Cluster- Nummer	$X_1$	$X_2$	$X_3$	$X_4$	$t$	$n$	<i>R41</i>		<i>Rest</i>	
							abs.	(%)	abs.	(%)
2	0	0	0	0	4	31	0	(0%)	31	(100%)

Demnach ist in dem ersten dieser beiden Cluster die Rate der reizenden Substanzen (Klasse *R41*) gleich 52%, im zweiten Cluster beträgt sie 0%. Der erste Cluster ( $n = 85$  Substanzen) ist dadurch charakterisierbar, daß *mindestens einer der 4 Prädiktoren den Wert "1" haben muß*. Dies wird in der schematischen Darstellung dadurch ausgedrückt, daß das "Muster" (1,1,1,1) für die 4 Prädiktoren ( $X_1, X_2, X_3, X_4$ ) an mindestens  $t = 1$  Stellen "getroffen" werden muß. Inhaltlich bedeutet dieses, daß mindestens einer der 4 Faktoren "vorhanden" oder "positiv" sein muß, damit ein Stoff in diese Konstellation gehört.

Umgekehrt ist der zweite Cluster (mit  $n = 31$  Substanzen) dadurch charakterisiert, daß *alle 4 Faktoren gleich 0 sind* ( $t = 4$  "Treffer" des Musters (0,0,0,0) der 4 Prädiktoren ( $X_1, X_2, X_3, X_4$ )).

Alle 41 reizenden Substanzen sind nach dieser Aufteilung also im ersten Cluster. Durch die Zuordnungsregel

Klassifiziere eine Substanz als "reizend", wenn mindestens einer der 4 Faktoren  $X_1$  bis  $X_4$  positiv ist; stufe sie als "nicht reizend" ein, wenn alle 4 Faktoren  $X_1$  bis  $X_4$  null sind.

erhält man daher eine Klassifikation mit einer Sensitivität von 100% und einer Spezifität von  $100 \times \frac{31}{31+41} = 43.1\%$  (jeweils in der Lernstichprobe!).

Die hier ausgewählte Partition ist also diejenige mit der *höchsten Sensitivität*. In der nachfolgenden Tabelle sind aus allen von *CBR* gefundenen Partitionen diejenigen mit der höchsten Sensitivität angegeben. Die zugehörigen Angaben beziehen sich jeweils auf denjenigen Cluster der Partition, dessen Substanzen in der Klassifikationsregel als "reizend" eingestuft würden:

Cluster mit niedrigem Anteil reizender (R41)- Substanzen						
Cluster-Nr.	Anzahl "nicht-reizender" Substanzen im Cluster		Anzahl "reizender" Substanzen im Cluster		Anzahl Substanzen insgesamt im Cluster	
				%-Satz "reizender" Substanzen	Sensitivität (%)	Spezi- fität (%)
1	31	0	31	0.0	100.0	43.1
2	37	2	39	5.1	95.5	51.4
3	42	6	48	12.5	86.4	58.3
4	52	6	58	10.3	86.4	72.2
5	53	8	61	13.1	81.8	73.7
6	58	8	66	12.0	81.8	80.6
7	58	9	67	13.4	79.5	80.6
8	57	10	67	14.9	77.2	79.2
9	58	12	70	17.4	72.7	80.6
10	63	12	75	16.0	72.7	87.5
11	63	13	76	17.1	70.5	87.5
12	65	14	79	17.7	68.2	90.3
13	64	15	79	19.0	65.9	88.9
14	66	16	82	19.5	63.6	91.7

Beim Vergleich zu den Ergebnissen von *CART*, dargestellt in der ROC-Kurve (Abb. 1), ist zu beachten, daß die dortigen Ergebnisse kreuzvalidiert und damit unverzerrt sind, während in dieser Tabelle bei den Angaben zu Sensitivität und Spezifität mit einer gewissen Überschätzung gerechnet werden muß. Dennoch geben diese Daten deutliche Hinweise darauf, daß es prinzipiell möglich ist, Klassifikationsregeln mit sehr hoher Sensitivität zu konstruieren, daß dieses jedoch (wie zu erwarten) nur unter Verzicht auf eine hohe Spezifität zu erreichen ist. Für das (im Rahmen von *CBR*) extremste Beispiel wurde die Konstruktionsregel oben angegeben.

Abschließend soll noch diejenige 2-er Partition dargestellt werden, die von *CBR* nach dem *P*-Wert-Kriterium als optimal gefunden wird:

Cluster- Nummer	$X_1$	$X_2$	$X_3$	$X_4$	$t$	$T_B$	$n$	$R41$		$Rest$	
								abs.	(%)	abs.	(%)
1	1	.	.	.	1	1	25	24	(96%)	1	(4%)
	.	1	1	.	2						

Cluster- Nummer	$X_1$	$X_2$	$X_3$	$X_4$	$t$	$T_B$	$n$	$R41$		$Rest$	
								abs.	(%)	abs.	(%)
2	0	.	.	.	1	2	91	20	(22%)	71	(78%)
	.	0	0	.	1						

Der erste Cluster dieser Partition besteht aus  $n = 25$  Substanzen, von denen 24 (= 96%) im *in vivo*-Versuch als reizend klassifiziert wurden. Im zweiten Cluster mit  $n = 91$  Substanzen waren 71 (= 17%) als *nicht-reizend* eingestuft worden.

Beide Cluster sind durch *zwei* Muster von Werten der Prädiktoren zu charakterisieren. Im Cluster 1 sind es die Muster  $(X_1) = (1)$  mit mindestens  $t = 1$  "Treffern", und  $(X_2, X_3) = (1, 1)$  mit mindestens  $t = 2$  "Treffern". Von diesen zwei Bedingungen müssen mindestens  $T_B = 1$  erfüllt sein.

Der Cluster 2 besteht aus allen anderen Substanzen. Sie sind also dadurch charakterisiert, daß die Bedingung  $X_1 = 0$  erfüllt ist (das Muster (0) für den Faktor  $(X_1)$  hat mindestens  $t = 1$  "Treffer"), und daß gleichzeitig die Bedingung "mindestens  $t = 1$  Treffer der Prädiktoren  $(X_2, X_3)$  im Muster (0,0)" erfüllt ist (von den beiden aufgeführten Bedingungen müssen mindestens  $T_B = 2$ , also beide, erfüllt sein). In anderen Worten:

Die Klasse 1 mit einem sehr hohen Anteil *reizender* Substanzen wird aus allen denjenigen Stoffen gebildet, die bei den *in-vitro* Ergebnissen entweder im ersten Faktor  $X_1$  oder in den beiden Faktoren  $X_2$  und  $X_3$  (oder in allen drei Faktoren  $X_1, X_2, X_3$ ) positiv sind.

Die Klasse 2 mit einem sehr hohen Anteil *nicht-reizender* Substanzen wird aus allen denjenigen Stoffen gebildet, die bei den *in-vitro* Ergebnissen sowohl im ersten Faktor  $X_1$  als auch in mindestens einem der beiden weiteren Faktoren  $X_2$  und  $X_3$  null sind.

Der Faktor  $X_4$  spielt bei dieser Klassifikation keine Rolle. Das bedeutet, daß eine Partition mit einem größeren  $\chi^2$ -Wert (einem kleineren  $P$ -Wert) der zugehörigen Vierfelder-Tafel auch durch die Einbeziehung des Faktors  $X_4$  nicht erreichbar ist. Der (minimale)  $P$ -Wert der hier dargestellten Partition beträgt  $1.42 \times 10^{-11}$ .

### 3.2.2 Partition in drei Cluster

Eine Partition der Menge aller möglichen Wertekonstellationen der Faktoren  $X_1$  bis  $X_4$  in *drei* Klassen wurde aufgrund der folgenden Überlegung durchgeführt:

Unter der Vorgabe einer 3-er Partition sucht *CBR* nach 3 disjunkten Clustern, die sich bezüglich der Rate der in ihr enthaltenen *R41*-Substanzen (gemessen am  $\chi^2$ -Wert der  $2 \times 3$ -Tabelle) maximal unterscheiden. Bei Anwendung dieses Verfahrens wird man also erwarten, 2 Cluster mit sehr hohem bzw. sehr niedrigem Anteil reizender Substanzen zu erhalten und als drittes einen Cluster, in dem reizende und nicht-reizende Substanzen stark gemischt sind. Eine solche Partition könnte als *partielle Klassifikation* benutzt werden: Substanzen, die in eine der beiden extremen Konstellationen fallen, werden als "reizend" bzw. als "nicht-reizend" klassifiziert. Für die Klassifikation

der übrigen Substanzen sind weitere Informationen und Kriterien einzuholen.

Zur Bewertung verschiedener solcher 3-er Partitionen ist es in Hinblick auf das Ziel der partiellen Klassifikation sinnvoll, über das von *CBR* benutzte Kriterium des minimalen *P*-Wertes hinaus noch weitere spezielle Kriterien zu definieren. Dabei sollten sowohl die Raten der richtig klassifizierten Fälle als auch der Anteil gar nicht klassifizierten Substanzen eingehen:

Sei  $(C_+, C_0, C_-)$  eine Partition der Menge aller möglichen Wertekombinationen der Faktoren  $X_1, X_2, X_3, X_4$  mit folgender Klassifikationsregel:

Substanzen, deren in vitro Ergebnisse nach  $C_+$  fallen, werden als "reizend" eingestuft, die  $C_-$ -Substanzen als "nicht-reizend" und die  $C_0$ -Substanzen bleiben unklassifiziert. Dann bezeichne

$$p_{+|+} = P(\text{Substanz in vivo} = \text{"R41"}) \mid \text{Substanz in vitro} \in C_+$$

die Wahrscheinlichkeit dafür, daß eine aufgrund der in vitro Ergebnisse als "reizend" klassifizierte Substanz auch im in vivo Versuch als "R41" klassifiziert wird.  $p_{+|+}$  wird auch der *positive prädiktive Wert* genannt. Analog sei

$$p_{-|-} = P(\text{Substanz in vivo} = \text{"nicht R41"}) \mid \text{Substanz in vitro} \in C_-$$

die Wahrscheinlichkeit dafür, daß eine aufgrund der in vitro Ergebnisse als "nicht-reizend" klassifizierte Substanz auch im in vivo Versuch als "nicht-R41" klassifiziert wird.  $p_{-|-}$  wird auch der *negative prädiktive Wert* genannt. Als *prädiktiver Wert* wird im folgenden das gewichtete Mittel aus positivem und negativem prädiktiven Wert definiert, wobei die Gewichtung proportional zur Anzahl von Substanzen  $n_+$  in  $C_+$  bzw.  $n_-$  in  $C_-$  in der Gesamtpopulation gewählt wird:

$$\text{Prädiktiver Wert } PW = \frac{p_{+|+} n_+ + p_{-|-} n_-}{n_+ + n_-}$$

Weiterhin bezeichne, wenn mit  $n_0$  die Anzahl der Fälle in  $C_0$  bezeichnet wird,

$$r_0 = \frac{n_0}{n_+ + n_- + n_0}$$

die Rate der nicht-klassifizierten Substanzen.

Eine gute partielle Klassifikation sollte dann einen möglichst hohen prädiktiven Wert *PW* bei gleichzeitig niedriger Rate  $r_0$  haben. Tatsächlich ist üblicherweise aber ein hoher prädiktiver Wert mit einer *hohen* Rate  $r_0$  verbunden und umgekehrt eine niedrige Rate  $r_0$  nicht klassifizierter Fälle mit einem niedrigen prädiktiven Wert. Zur Illustration ist in der folgenden Abbildung für die (nach dem *P*-Wert Kriterium von *CBR*) 100 besten 3-er Partitionen jeweils die Rate  $r_0$  und der prädiktive Wert gegeneinander aufgetragen:

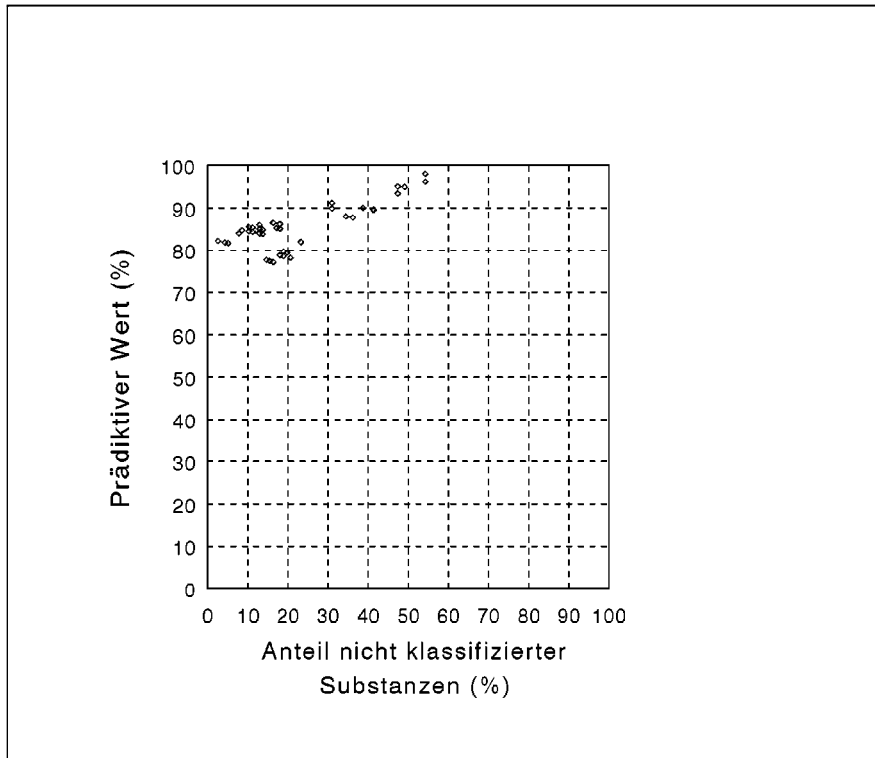


Abb. 7: Prädiktiver Wert und Anteil nicht klassifizierter Substanzen für verschiedene partielle Klassifikationsregeln

Aus der Abbildung geht hervor, daß der höchste prädiktive Wert von 98% nur bei einem Anteil von  $r_0 = 54\%$  unklassifizierter Substanzen erreicht werden kann. Auf der anderen Seite ist bei einem Anteil von 3% unklassifizierter Fälle nur noch ein prädiktiver Wert von 82% erreichbar.

Die Partition mit dem höchsten prädiktiven Wert kann wie folgt beschrieben werden:

Cluster (Partition)	$X_1$	$X_2$	$X_3$	$X_4$	$t$	$T_B$	$n$	$R_{41}$		$Rest$	
								abs.	(%)	abs.	(%)
$C_+$	1	.	.	.	1	1	22	21	(95%)	1	(5%)

Cluster (Partition)	$X_1$	$X_2$	$X_3$	$X_4$	$t$	$T_B$	$n$	$R_{41}$		$Rest$	
								abs.	(%)	abs.	(%)
$C_0$	0	.	.	.	1	2	63	23	(37%)	40	(63%)
	.	1	1	1	1						

Cluster (Partition)	$X_1$	$X_2$	$X_3$	$X_4$	$t$	$T_B$	$n$	$R_{41}$		$Rest$	
								abs.	(%)	abs.	(%)
$C_-$	0	0	0	0	4	1	31	0	(0%)	31	(100%)

$C_+$  kann also einfach dadurch charakterisiert werden, daß der erste Faktor ( $X_1$ ) positiv ist (daß also der Mittelwert der Reaktionszeit bei Reizindex 10% kleiner oder gleich 109 sec. ist). Eine Einordnung in  $C_-$  erfolgt dann, wenn alle vier Faktoren  $X_1$  bis  $X_4$  gleich null sind. Zu den nicht-klassifizierten gehören alle übrigen Substanzen, die also im ersten Faktor  $X_1$  gleich

0, aber mindestens in einem der weiteren Faktoren  $X_2, X_3, X_4$  positiv sind.

Die nach dem Standard-Kriterium von *CBR* bevorzugte Partition mit dem kleinsten  $P$ -Wert hat einen prädiktiven Wert von 86% bei einer Rate von  $r_0 = 13\%$  nicht klassifizierter Substanzen. Sie ist wie folgt zu beschreiben:

Cluster (Partition)	$X_1$	$X_2$	$X_3$	$X_4$	$t$	$T_B$	$n$	$R41$		$Rest$	
								abs.	(%)	abs.	(%)
$C_+$	1	1	.	.	1	2	22	22	(100%)	0	(0%)
	.	.	1	.	1						

Cluster (Partition)	$X_1$	$X_2$	$X_3$	$X_4$	$t$	$T_B$	$n$	$R41$		$Rest$	
								abs.	(%)	abs.	(%)
$C_0$	1	1	.	.	1	2	15	8	(53%)	7	(47%)
	.	.	0	.	1						

Cluster Partition)	$X_1$	$X_2$	$X_3$	$X_4$	$t$	$T_B$	$n$	$R41$		$Rest$	
								abs.	(%)	abs.	(%)
$C_-$	0	0	.	.	2	1	79	14	(18%)	65	(82%)

Der  $P$ -Wert dieser Partition beträgt  $7.56 \times 10^{-12}$ , ist also kleiner als der minimale  $P$ -Wert der 2-er Partition. Dies kann als Hinweis darauf gewertet werden, daß die Zusammenhangsstruktur zwischen den in vitro-Faktoren  $X_1$  bis  $X_4$  und der in vivo-Einstufung ( $R41$ ) mit der differenzierteren Klassifizierung in drei Partitionen besser als mit einer 2-er Partition erfaßt wird.

## 4 Schlußfolgerung

Es wurde gezeigt, in welcher Weise das Baumanalyseprogramm *CART* zur Diskrimination zwischen  $R41$ - ("reizenden") Substanzen und "nicht-reizenden" Substanzen unter Verwendung der in vitro-Ergebnisse eingesetzt werden kann. Insbesondere konnten dabei unter Anwendung der Kreuzvalidierung und unter Variation der Fehlklassifikationskosten im ROC Diagramm die Möglichkeiten und Grenzen der angewendeten in vitro-Tests in Bezug auf die Klassifikation der in vivo-Tests aufgezeigt werden.

Als Vorteil von *CART* gegenüber klassischen Verfahren der Diskriminanzanalyse kann es angesehen werden, daß die Zuordnungsregel *sequentiell* ist, so daß also die Anwendung weiterer in vitro-Tests nur bei bestimmten Ergebniskonstellationen der vorangehenden Tests erforderlich ist. Zu beachten ist weiterhin, daß fehlende, nicht erhältliche oder nicht verwendbare Ergebnisse von in vitro-Tests bis zu einem gewissen Umfang durch andere Meßgrößen (Surrogate) ersetzt werden können.

Es wurde auch gezeigt, daß in dem vorliegenden Datensatz Wechselwirkungen zwischen den einzelnen in vitro-Testergebnissen untereinander in Beziehung zu ihrer Diskriminationsfähigkeit vorhanden sind und damit eine Situation vorliegt, die von *CART* (im Gegensatz zur linearen Diskriminanzanalyse) flexibel erfaßt werden kann.

Ähnlich wie *CART* ist auch das Programm *CBR* in der Lage, etwaige Wechselwirkungseffekte zu erfassen. Es zeigte sich, daß eine Kombination von *CART* und *CBR* dazu benutzt werden kann,

nach Klassifikationsregeln zu suchen, die den speziellen Anforderungen der vorliegenden Fragestellung genügen. Hierzu gehört insbesondere die Forderung nach einer hohen Sensitivität der Zuordnungsregel, also nach einer sehr hohen Wahrscheinlichkeit dafür, daß eine als "R41" einzustufende Substanz auch durch die in vitro-Tests als "reizend" erkannt wird.

Als eine allgemeine *inhaltliche* Schlußfolgerung aus den Ergebnissen dieser Analyse kann angesehen werden, daß eine Klassifikationsregel mit gleichzeitig hoher Sensitivität und hoher Spezifität (z.B. beide Werte deutlich über 80%) bei realistischer Einschätzung, d.h. bei Betrachtung der kreuzvalidierten Ergebnisse, *nicht* erreichbar ist. Es sollte daher in diesem Zusammenhang auch die Möglichkeit der *partiellen* Klassifikation in Betracht gezogen werden. Dabei wird auf eine Klassifikation *aller* Substanzen auf der alleinigen Basis der in vitro-Testergebnisse verzichtet. Es sollen aber diejenigen Substanzen selektiert werden, von denen man aufgrund der in vitro-Tests mit sehr hoher Wahrscheinlichkeit ihre potentielle in vivo-Klassifikation richtig vorhersagen kann (die Zuordnung sollte für die danach selektierten Substanzen einen hohen *prädiktiven Wert* haben).

Es wurde mit Hilfe von *CBR* – wiederum in Verbindung mit *CART* – gezeigt, welche Kombinationen von prädiktivem Wert und Anteil nicht-klassifizierbarer Fälle in dem vorliegenden Datensatz möglich sind. Gleichzeitig wurden spezielle partielle Klassifikationsregeln explizit angegeben.

## References

- [1] ANDERSON.T.W. *An introduction to multivariate statistical analysis*. New York, (1958).
- [2] L. BREIMAN, J.H. FRIEDMAN, R.A. OLSHEN, and C.J. STONE. *Classification and Regression Trees*. Belmont, California, (1984).
- [3] H. HECKER und P. WÜBBELT. Klassifikation nach Mustern von Prognosefaktoren und Ausprägungen von Responsevariablen. In S. Schach and G. Trenkler, editors, *Data Analysis and Statistical Inference. Festschrift in Honour of Prof.Dr.F.Eicker*, pages 259–275. Eul, Bergisch Gladbach, Köln, (1992).
- [4] H. HECKER and P. WÜBBELT. Clustering By Response: CBR. *Comput Stat Data Anal.* 24, 193–215 (1997)
- [5] H. HECKER and P. WÜBBELT. *Clustering By Response: CBR – Einführung und Programmbeschreibung*. Medizinische Hochschule Hannover (1994).
- [6] S. GLASER. *Substanzklassifikation in Toxizitätsklassen mit Verfahren der Diskriminanzanalyse*. Statusseminar "Biometrische Methoden zur Planung, Auswertung und Validierung von in vitro-Verfahren als Ersatz für Tierversuche in der Toxikologie" (1996).
- [7] J.N. MORGAN and J.A. SONQUIST. Problems in the analysis of survey data, and a proposal. *J. Amer. Statist. Assoc.* 58: 415-434 (1963).