SAS Makro zur CART-Analyse am Beispiel eines dermato-epidemiologischen Datensatzes

Martina Malzer
IMBE Universität Erlangen-Nürnberg
Waldstr. 6
91054 Erlangen
Martina.Malzer@imbe.med.uni-erlangen.de

Annette Pfahlberg
IMBE Universität Erlangen-Nürnberg
Waldstr. 6
91054 Erlangen
Annette.Pfahlberg@imbe.med.uni-erlangen.de

Wolfgang Uter
IMBE Universität Erlangen-Nürnberg
Waldstr. 6
91054 Erlangen
Wolfgang.Uter@imbe.med.uni-erlangen.de

Janice Hegewald
Institut und Poliklinik für Arbeits- und Sozialmedizin, Technische Universität Dresden
Fetscherstr. 74
01307 Dresden
janice.hegewald@mailbox.tu-dresden.de

Olaf Gefeller
IMBE Universität Erlangen-Nürnberg
Waldstr. 6
91054 Erlangen
Olaf.Gefeller@imbe.med.uni-erlangen.de

Zusammenfassung

Die UV-Lichtempfindlichkeit der Haut steht direkt im Zusammenhang mit der Augenfarbe, Haarfarbe, den Sommersprossen, der Anzahl der Muttermale an beiden Armen und der Hauthelligkeit (in unserem Beispiel gemessen mittels Reflektometer).
Die Empfindlichkeit der Haut kann in verschiedene Hauttypen eingeteilt werden (Fitzpatrick-Hauttyp). Allerdings ist der Hauttyp in einem selbst auszufüllenden Fragebogen schwer ermittelbar. Aus diesem Grund soll versucht werden, den Lichthauttyp auf der Basis anderer, reliabel messbarer konstitutioneller Merkmale zu bestimmen.
Es wurde ein automatisiertes SAS Makro verwendet, welches obige Merkmale mittels p-Wert basiertem CART geeignet miteinander kombiniert. Für die Realisierung einer p-Wert basierten CART-Analyse innerhalb des automatisierten SAS Makros steht in SAS die FREQ-Prozedur zur Verfügung, welche die p-Werte ausgibt.
In einem nächsten Schritt werden die p-Werte der durchgeführten Tests miteinander verglichen, um so den kleinsten p-Wert für einen Split (Aufteilung) erhalten zu können.
Daraus resultiert EIN binärer Klassifikationsbaum, in dem die oben genannten Merkmale bestmöglich miteinander kombiniert dargestellt werden, um die Zuordnung in einen Haut-

typ erleichtern zu können. Das Verfahren wird anhand eines Bespieldatensatzes über eine Befragung zum Thema Sonnenschutz/Verhalten in der Sonne bei jungen Erwachsenen illustriert.

Schlüsselwörter: p-Wert basiertes CART, SAS Makro, Klassifikation, Baumverfahren

1 Hintergrund / Fragestellung

Im Rahmen einer epidemiologischen Befragung zum Thema Sonnenschutz/Verhalten in der Sonne wurden N=1568 Studenten anhand eines selbstauszufüllenden Fragebogen identifiziert. Die Teilnehmer bestimmten u. a. ihren Konstitutionstyp (Haarfarbe, Augenfarbe und Sommersprossen), die Anzahl der Muttermale beider Arme und die Hauthelligkeit mithilfe eines sogenannten Reflektometers: *Reflectometer RM 100, Courage und Khazaka, Köln*. Es misst den Melaningehalt in der Haut (→ Je größer der Wert, desto heller ist die Haut).
Des Weiteren erhoben die Befragten ihre UV-Lichtempfindlichkeit nach dem Fragenkomplex zur Einteilung in die Hautempfindlichkeitstypen nach Fitzpatrick [1].
Die Fragestellung im Fragebogen zur Selbstermittlung des Hauttyps nach Fitzpatrick lautet: „Wie reagiert Ihre Haut, wenn sie um die Mittagszeit am ersten sonnigen Sommertag eine halbe Stunde ungeschützt (ohne vorheriges Eincremen) der Sonne ausgesetzt wird?“

- Hauttyp I: „Immer rot, nie braun“
- Hauttyp II: „Immer rot, manchmal Bräunung“
- Hauttyp III: „Manchmal rot, immer Bräunung“
- Hauttyp IV: „Nie rot, immer Bräunung“

Allerdings stellte sich die Selbstklassifikation der Teilnehmer in einen der Fitzpatrick-Hauttypen als schwierig heraus. Deshalb soll versucht werden, den Lichthauttyp auf der Basis anderer, reliabel messbarer konstitutioneller Merkmale zu bestimmen.
Die UV-Lichtempfindlichkeit der Haut steht direkt im Zusammenhang mit der Augenfarbe, Haarfarbe, Sommersprossen, Anzahl der Muttermale und der Helligkeit der Haut [2].

Somit entstand die Idee, alle mit der UV-Lichtempfindlichkeit assoziierten Merkmale wie Augenfarbe, Haarfarbe, Sommersprossen, Anzahl der Muttermale und die Reflektometer-Messwerte kombiniert zu betrachten mithilfe eines geeigneten Analyseverfahrens. Durch die Analyse soll eine bessere Klassifikation der UV-Lichtempfindlichkeit erzielt werden.

2 Methode

Als geeignetes kombinierendes Analyseverfahren wurde eine p-Wert basierte CART-Analyse gewählt, die innerhalb des automatisierten SAS Makros „chiklass“ bereits implementiert ist. Aus dem Verfahren (SAS Makro) resultiert EIN binärer Klassifikations-

baum, in dem alle Merkmale (Prädiktoren) bestmöglich miteinander kombiniert dargestellt werden, um die Zuordnung in einen Hauttyp erleichtern zu können.

2.1 p-Wert basierte CART-Analyse (CART)

Der Begriff CART steht für „Classification And Regression Trees“ und stellt seit 1984 einen bedeutenden und leistungsfähigen Algorithmus zur Entscheidungsfindung in Form von binären Baumstrukturen dar [3]. Dabei werden für eine Entscheidung pro Verzweigung (Split) höchstens zwei Äste gebildet. Eine Entscheidung für die Einteilung in binäre Äste wird durch einfache Fragestellungen getroffen, die meist mit „ja“ oder „nein“ zu beantworten sind. Ein CART beschreibt also eine Folge von Entscheidungsfragen, die in der Kombination betrachtet werden. Zuletzt soll anhand der Entscheidungsfragen ein Ergebnis aus dem Baum erhalten werden, nach dem entsprechend entschieden werden kann (Klassifikation).
Beim CART-Verfahren nach [3] werden vor der Analyse die Daten in eine Lern- und in eine Teststichprobe aufgeteilt zum Finden des optimalen Baumes. Dabei bildet die Lernstichprobe den Teil der Datenmenge zur Findung des Klassifizierers. Die Teststichprobe beschreibt den Teil der Datenmenge, der für die Schätzung der absoluten Fehlerrate (Test) für eine Missklassifikation behalten wird – vgl. [3].

Die p-Wert basierte CART-Analyse ist ein testbasiertes Verfahren und kein modellbasiertes Verfahren wie in [3]. Die p-Werte für das Finden der Splits (Verzweigungen) stammen aus dem χ^2-Test bzw. dem exakten Test nach Fisher [4], wobei immer der kleinste p-Wert aus dem Vergleich der p-Werte aus der Teststatistik gewählt wird. Ein festgesetztes Signifikanzniveau α darf aber dabei nicht überschritten werden. Es gilt: $p \leq \alpha$.
Letztendlich soll nur EIN Klassifikationsbaum unter Kombination der verschiedenen Merkmale (Prädiktoren) generiert werden.
Um einen Klassifikationsbaum mit binärer Struktur zu gewährleisten, sollen sowohl die kategorielle Zielgröße als auch die miteinander geeignet zu kombinierenden Merkmale bzw. Prädiktoren dichotom in die Analyse eingehen. Alle Größen sollen somit je 2 Ausprägungen (Kategorien) besitzen.
Zielgröße für unsere Analyse ist die Hautempfindlichkeit und als Prädiktoren werden Augenfarbe, Haarfarbe, Sommersprossen, Anzahl der Muttermale an beiden Armen und der Reflektometer-Messwert durch p-Wert basiertes CART entsprechend miteinander kombiniert betrachtet.

2.2 Dichotomisierung der Analysedaten

Eine Dichotomisierung ist immer dann notwendig, wenn eine zu testende Größe binär sein soll. Ein geeigneter „Schwellenwert“ soll dabei die beiden entstehenden Gruppen der Testgröße optimal voneinander trennen.
Mehr als zweistufige kategorielle bzw. stetige Merkmale können z. B. anhand ihrer Verteilung in zweistufige Merkmale überführt werden. Die daraus folgende Zweigrup-

pen-Aufteilung durch Zusammenfassung der einzelnen Ausprägungen (Kategorien) eines Merkmals kann z. B nach dem Median erfolgen (sodass in jeder Gruppe ca. 50% der Probanden sind). So ist eine Ausgewogenheit in den neu entstandenen Gruppen gewährleistet.

Beispiele

Das sechsstufige Merkmal Augenfarbe unserer Daten mit den Ausprägungen: dunkelblau, hellblau/grau, grün, grün/braun, hellbraun, dunkelbraun
wurde zusammengefasst in:

- Gruppe 1: dunkelblau, hellblau-grau, grün (51% der Verteilung)
- Gruppe 2: grün-braun, hellbraun, dunkelbraun (49% der Verteilung)

Des Weiteren besteht gerade bei mehr als zweistufigen kategoriellen Merkmalen die Möglichkeit, deren Kategorien nach inhaltlichen Aspekten in zwei Gruppen aufzuteilen. Dies wurde z. B. für die vierstufige (Ziel-)Variable für den Fitzpatrick-Hauttyp durchgeführt:

- Gruppe 1: Fitzpatrick-Hauttyp I + II (UV-lichtempfindlichere Hauttypen)
- Gruppe 2: Fitzpatrick-Hauttyp III + IV (weniger UV-lichtempfindliche Hauttypen)

Bei stetigen Variablen wie die Anzahl der Muttermale an beiden Armen oder der Reflektometer-Messwert konnte der Schwellenwert (Cutpoint oder Cut-off-Point) anhand von ROC-Kurven (Receiver Operating Characteristic - Kurve) bestimmt werden [5].
Folgende Cutpoints konnten für unsere Daten durch ROC-Kurven ermittelt werden:

- Anzahl der Muttermale an beiden Armen: Cutpoint bei Anzahl „13“
- Reflektometer-Messwert: Cutpoint bei Messwert „40“

Die Voraussetzung, um das SAS Makro „chiklass“ nutzen zu können, ist die vorherige Dichotomisierung aller Variablen (Zielgröße und Prädiktoren).
Definition aller nun dichotomisierten Daten mit Kodierungen für das SAS Makro:

- Zielgröße: Hautempfindlichkeit:
  - 1 = UV-lichtempfindlicher Hauttyp (Fitzpatrick I + II)
  - 0 = weniger UV-lichtempfindlicher Hauttyp (Fitzpatrick III + IV)
- Prädiktor 1: Sommersprossen:
  - 1 = Sommersprossen
  - 0 = keine Sommersprossen
- Prädiktor 2: Reflektometer (Cutpoint: „40“, s. 2.2.1):
  - 1 = Messwert > 40
  - 0 = Messwert ≤ 40
- Prädiktor 3: Haarfarbe:
  - 1 = rote / blonde Haare
  - 0 = braune / schwarze Haare

- Prädiktor 4: Anzahl der Muttermale (Cutpoint: „13“, s. 2.2.1):

  - 1 = Anzahl Muttermale > 13
  - 0 = Anzahl Muttermale ≤ 13

- Prädiktor 5: Augenfarbe:
  - 1 = blaue / graue / grüne Augen
  - 0 = braune Augen

Als weitere Voraussetzung für die Verwendung des SAS Makros sollen als Kodierungen für die neuen zweistufigen Variablen (Größen) nur die Werte 0 und 1 verwendet werden (s. oben, Definition mit Kodierungen).

3 Beschreibung des SAS Makros *chiklass*

Für die Analyse der in 2.1 beschriebenen Methodik steht das SAS Makro „chiklass“ zur Verfügung. Für das Finden der Splits zur Generierung des Klassifikationsbaums werden im SAS Makro die p-Werte aus dem exakten Test nach Fisher der SAS-Prozedur FREQ verwendet. Der **kleinste** p-Wert ist dabei optimales Splitkriterium. Jedoch darf dieser das im SAS Makro gewählte Signifikanzniveau α nicht überschreiten, sonst erfolgt an der Stelle keine weitere Verzweigung für den Klassifikationsbaum (Default für das Signifikanzniveau: `alpha = 0.05`).

Als ein weiteres Stoppkriterium für den Baum ist $n < 10$ implementiert, wobei n die Anzahl der Getesteten beschreibt (also die „Datenmenge“, für die der exakte Fisher-Test gerechnet wird).

Beschreibung des Aufrufs von „chiklass“:

```
datei=        aktuell zu verwendender  Datensatz für die Auswertung

markervar=  Liste aller zu verwendenden  Prädiktoren

ziel=          abhängige Variable (Zielvariable)

alpha=0.05    Signifikanzniveau, Default: 5%; frei wählbar

kzustand=1  Darstellung im Output
              Default: 1 nach Fälle  Optional 0 nach Nicht-Fälle

output=1    Ergebnisse werden in Tabellenform in ein RTF oder PDF
            geschrieben

grafik=1     Ergebnisse werden durch eine Baumgrafik
             in einer RTF- oder PDF-Datei dargestellt
            (JEDE andere Belegung der Var. ergibt KEIN Output!)
            Formate:  A3 bei > 6 Var, A4 bei <=6 Var; A3: Zeichnung
            ist begrenzt auf 7 Ebenen wg. Übersichtlichkeit!

 outformat=    Dateiformat für Tabelle/Baumgrafik: rtf ODER pdf

 out=      Ausgabe-Pfad wählbar für Ergebnistabelle UND
           Baumgrafik (wäre für beides gültig!)
/*************************************************************************/
 AUFRUF:
 %chiklass(datei=saslibrary.dateiname || datei = dateiname,
            markervar=Variable1  Variable2 ...   ,
            ziel=Zielvariable, output=1, grafik=1, outformat=rtf,
            out=Laufwerk:\Verzeichnis\);
/*************************************************************************/
```

Abbildung 1: Allgemeine Makrobeschreibung

Aufruf des Makros „chiklass“ mit unseren Daten:

```
%chiklass(datei=lib.daten_fuer_macro,
          markervar=s2_naevi_gesamt2mm_bin reflektometer_bin
          sprossen augenfarbe haarfarbe,
          ziel=helligkeit, alpha=0.05, kzustand=1, output=1,
          grafik=1, outformat=rtf,
          out=K:\KSFE2012\Outputs);
```

(Für `alpha`, `kzustand`, `output`, `grafik`, `outformat` werden Defaulteinstellungen übernommen.)

Pro Split wird jeder Prädiktor in Variablenliste `markervar` innerhalb der FREQ-Prozedur verglichen mit der Zielvariable `helligkeit` (Hauthelligkeit). Da die Einstellung `kzustand=1` beibehalten wird (Beschreibung s. Abb. 1), wird die UV-lichtempfindlichere Gruppe (`helligkeit=1`) als exponierte Gruppe betrachtet, sowohl in der Analyse als auch im späteren Output (Ergebnisse und/oder Baumgrafik).

Beispiel für den Output des 1. Splits aus der PROC FREQ (Sommersprossen):

Table of sprossen by helligkeit

sprossen(Sommersprossen)
helligkeit(UV-Lichtempfindlichkeit)

Frequency Percent Row Pct Col Pct	1	0	Total
1	338 21.56 48.49 56.81	359 22.90 51.51 36.90	697 44.45
0	257 16.39 29.51 43.19	614 39.16 70.49 63.10	871 55.55
Total	595 37.95	973 62.05	1568 100.00

Statistic	DF	Value	Prob
Chi-Square	1	59.2776	<.0001
Likelihood Ratio Chi-Square	1	59.3289	<.0001
Continuity Adj. Chi-Square	1	58.4740	<.0001
Mantel-Haenszel Chi-Square	1	59.2398	<.0001
Phi Coefficient		0.1944	
Contingency Coefficient		0.1909	
Cramer's V		0.1944	

Fisher's Exact Test

Cell (1,1) Frequency (F)	338
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	1.006E-14
Table Probability (P)	5.619E-15
Two-sided Pr <= P	1.669E-14

Sample Size = 1568

Abbildung 2: Output aus Proc Freq für den 1. Split

Eine gefundene Prädiktorvariable aus der Variablenliste (s. Makroaufruf) mit ihrem kleinsten p-Wert kann pro Teilbaum nur einmal berücksichtigt werden. Schließlich soll pro Variable und Teilbaum nur ein Split erfolgen. Für den ersten gefundenen Prädiktor des ersten Splits aus Abb. 2 bedeutet dies: Variable `sprossen` (Sommersprossen) wird für alle weiteren PROC FREQs zur Generierung der anderen Teilbäume ausgeschlossen. Die Funktion des Makros „chiklass“ soll durch folgendes Ablaufdiagramm verdeutlicht werden:

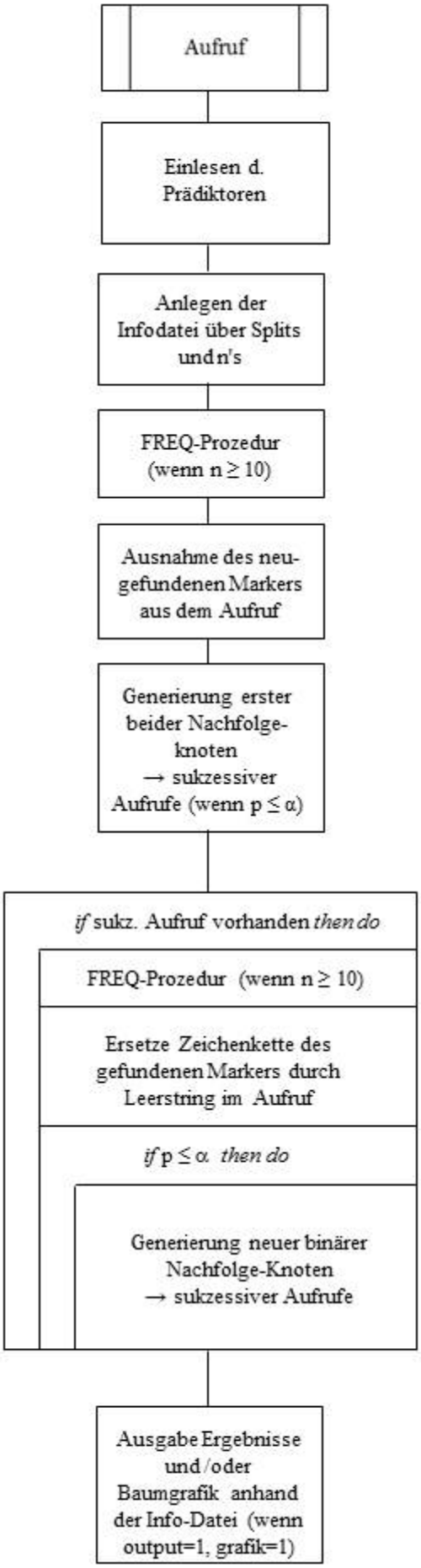

Abbildung 3: Flussdiagramm des SAS Makros

Nach Aufruf des Makros „chiklass“ werden alle Variablen der Variablenliste (im Aufruf mit Leerzeichen voneinander getrennt) an Makrovariablen übergeben. Anschließend wird zur Steuerung eine Info-Datei zu den Splits und den „*n*’s“ (Knoten) angelegt, welche später zur Orientierung und dem richtigen „Anspringen“ innerhalb der Knoten dient. Die maximale Größe des zu erwartenden resultierenden binären Klassifikationsbaums ist bekannt.
Die Generierung des ersten Splits wird im Programm aus didaktischen Gründen sequentiell durchlaufen. Wenn die Anzahl *n* als zu testende „Datenmenge“ die Grenze 10 nicht unterschreitet, werden mittels einer oder mehrerer PROC FREQs (abhängig von der Anzahl der in die Analyse eingehenden Prädiktoren) alle p-Werte berechnet. Sobald der minimale p-Wert automatisch ausgewählt ist, wird der Name des für den ersten Split gefundenen Prädiktors für den nächsten PROC FREQ-Durchlauf aus der aktuellen Variablenliste eliminiert. So kann eine Variable immer nur einmal pro Teilbaum in der Analyse berücksichtigt werden.
Die Information, welcher Prädiktor für einen Split schon einmal gefunden wurde und ob infolge „sukzessive Aufrufe“ (Voraufrufe) für nachfolgende Knoten zu notieren sind, wird wieder in die Info-Datei geschrieben. Die neu generierten sukzessiven Aufrufe der Nachfolgeknoten verweisen innerhalb der Info-Datei wieder auf die n’s, für welche die Teststatistik in einem nächsten Schritt erneut durchlaufen wird. So werden alle nachfolgenden Splits innerhalb einer Schleife durchlaufen (s. Abb. 3), solange die festgelegten Bedingungen ($p \leq \alpha$ oder $n \geq 10$) erfüllt sind. Die Durchläufe durch die sukzessiven Aufrufe werden kontrolliert durch einen sogenannten Pointer („Laufvariable“).
Am Ende der Analyse werden Informationen aus der Steuerungsdatei auch für das Erstellen der beiden Outputs (Ergebnistabelle und Baumgrafik) benötigt, wenn im Aufruf die entsprechenden Angaben durch den Nutzer erfolgt sind (`output=1`, `grafik=1`).

4 Ergebnisse/Interpretation

Im SAS Makro „chiklass“ kann der Anwender für seine Analyse sowohl die Ergebnisse detailliert in Tabellenform als auch eine zusätzliche Baumgrafik anfordern. Es ist auch nur eine Ausgabeform der Ergebnisse möglich. Beide Datei-Outputs werden programmtechnisch mithilfe von SAS ODS (SAS Output Delivery System) erstellt. Das Ausgabeformat der Dateien kann im Makro gewählt werden. Hier besteht die Option, entweder die beiden Outputs mit dem Aufruf-Statement `outformat=rtf` als RTF-Dateien oder mit `outformat=pdf` als PDF-Dateien erzeugen zu lassen.

Für die Analyseresultate unserer Daten werden die Ergebnistabelle (für eine detaillierte Übersicht der Ergebnisse) und die Baumgrafik (zur Darstellung des gefundenen binären Klassifikationsbaums) als RTF-Dateien angefordert (s. auch Makro-Aufruf).

4.1 Tabellarischer Output

Anhand des folgenden Tabellenauszugs aus ODS RTF unserer erhaltenen Resultate sollen die Notationen im Output und unsere wichtigsten Teilergebnisse interpretiert werden:

Ergebnisse

A

Gesamte Fälle	N	Anteil Fälle
595	1568	0.379

B

Knoten	p-Wert bei Split	Anzahl Fälle	Anzahl Getesteter	Anteil getesteter Fälle	Klassifikation
sprossen0	0.00000	257	871	0.295	**kein Zusammenhang**
sprossen1	0.00000	338	697	0.485	*Zusammenhang mit Zielvariable*
haarfarbe00	0.00000	142	617	0.230	**kein Zusammenhang**
haarfarbe01	0.00000	115	254	0.453	*Zusammenhang mit Zielvariable*
reflektometer_bin10	0.00000	161	407	0.396	*Zusammenhang mit Zielvariable*
reflektometer_bin11	0.00000	177	290	0.610	*Zusammenhang mit Zielvariable*

Abbildung 4: Auszug SAS ODS Output der Ergebnisse nach der Analyse

In unserer Analyse ist die exponierte Gruppe der Zielgröße „Hauthelligkeit“ die UV-lichtempfindlichere Gruppe mit `helligkeit=1`. Somit werden durch diese Gruppe die Fälle beschrieben mit n=595 (Anteil: 37.9%) von insgesamt N=1568 Befragten (s. Abb. 4, Tab. A). Für eine Klassifikation in die Gruppe der UV-lichtempfindlicheren Hauttypen wird der Anteil der Fälle aus Tabelle A als kritischer Wert mit 37.9% betrachtet.

„Sommersprossen“ ist das wichtigste Kriterium für die Klassifikation in die hautempfindlichere Gruppe, da dessen p-Wert auch zugleich als minimaler p-Wert im ersten Split gefunden wurde (vgl. dazu auch Abb. 2).
Der zweite Nachfolgeknoten `sprossen1` (s. Abb. 4, Tab. B, zweite Zeile) wurde programmintern zur Orientierung mit der Ordnungszahl 1 versehen, da nur diejenigen Beobachtungen enthalten sind, bei denen Sommersprossen vorhanden sind (`sprossen=1` für n=697, entspricht der Anzahl Getesteter). Somit haben n=338 (48.5%) der UV-lichtempfindlicheren Befragten Sommersprossen.
Für diesen Knoten gilt also die Klassifikation „Zusammenhang mit Zielvariable“ (hier: Zusammenhang Sommersprossen mit hoher UV-Lichtempfindlichkeit), da 48.5% größer als der kritische Wert 37.9% ist. Der Name des ersten Nachfolgeknotens `sprossen0` (s. Abb. 4, Tab. B, erste Zeile) ist mit der Ordnungszahl 0 versehen für die Beobachtungen ohne Sommersprossen (`sprossen=0` für n=871 getesteter Personen). 29.5% (n=257) der UV-lichtempfindlicheren Personen haben keine Sommersprossen. Es besteht kein Zusammenhang zwischen hoher UV-Lichtempfindlichkeit und „keine Sommersprossen“, da $29.5\% \leq 37.9\%$.

Der Reflektometer-Messwert ist das zweite Kriterium für die Klassifikation in die UV-lichtempfindlichere Gruppe: 61% (n=177) der UV-lichtempfindlicheren Befragten haben Sommersprossen und einen höheren Messwert (s. Abb. 4, Tab. B, sechste Zeile). Notation: `reflektometer_bin11` aus `sprossen=1` und `reflektometer_bin=1` (s. Abschnitt 3, Variablendefinitionen).

Die Haarfarbe ist ebenfalls ein Kriterium für die Eingruppierung in die UV-lichtempfindlichere Gruppe: 45.3% (n=115) der Befragten ohne Sommersprossen haben rote oder blonde Haare (s. Abb. 4, Tab. B, vierte Zeile). Notation: `haarfarbe01` aus `sprossen=0` und `haarfarbe=1`.

Die Ordnungszahlen werden je nach Belegung der „Nullen und Einsen" der Knotenvariablen (gefundene Prädiktoren) pro Split an den neuen Knotennamen angehängt. Deshalb ist es wichtig, vor der Analyse alle in das SAS Makro eingehenden zweistufigen Größen mit 0 oder 1 zu kodieren. Die Notationen aus 0 und 1 helfen ebenfalls bei der Interpretation.

4.2 Output Baumgrafik

Die Zeichnung des Klassifikationsbaums wird innerhalb des SAS Makros durch Pfeil- und Textelemente aus SAS ANNOTATE (SAS/GRAPH) gewährleistet. Das Papierformat (DIN A3 oder DIN A4) richtet sich nach der Anzahl der eingehenden Prädiktoren in der Variablenliste des Aufrufs (s. Abb. 1).
Bei einer Anzahl von bis zu sechs Prädiktoren erfolgt die Ausgabe in DIN A4, bei mehr als sechs Prädiktoren wird automatisch DIN A3 gewählt. Allerdings ist die Baumgrafik aus Gründen der Übersichtlichkeit auf sieben Ebenen beschränkt – im besten Fall würden bereits 2^7 (=128) Knoten auf der untersten (siebten) Ebene nebeneinander dargestellt werden.
Alle für die Grafik benötigten Informationen stammen wieder aus der Steuerungsdatei des Makros (s. Abb. 3), wie Knotennamen, Ausprägung 0 (in der späteren Zeichnung als − markiert) oder 1 (in der späteren Zeichnung als + markiert) des Knotens und die Anzahlen *n*. Die Grafikdatei für ANNOTATE enthält zusätzlich zu den oben genannten Informationen nur noch die im Makro berechneten Koordinaten für die grafische Darstellung der einzelnen Zeichnungselemente.
Des Weiteren werden vor der Generierung der Baumgrafik für die Namen der zu zeichnenden Prädiktorvariablen je dreistellige Alias-Namen vergeben, die durch den Nutzer nach automatischer Aufforderung eingegeben werden. Programmintern werden somit über die Variablennamen nur Labels gelegt. Diese Alias-Namen werden aus Gründen der besseren Übersichtlichkeit erzeugt. Im späteren RTF- oder PDF-Output werden die Variablennamen der Prädiktoren mitsamt ihrer zugehörigen Alias-Namen als Zusatzinformation automatisch über bzw. unter die Grafik geschrieben.

Für unsere Analysedaten bekommen wir folgenden Klassifikationsbaum als Output:

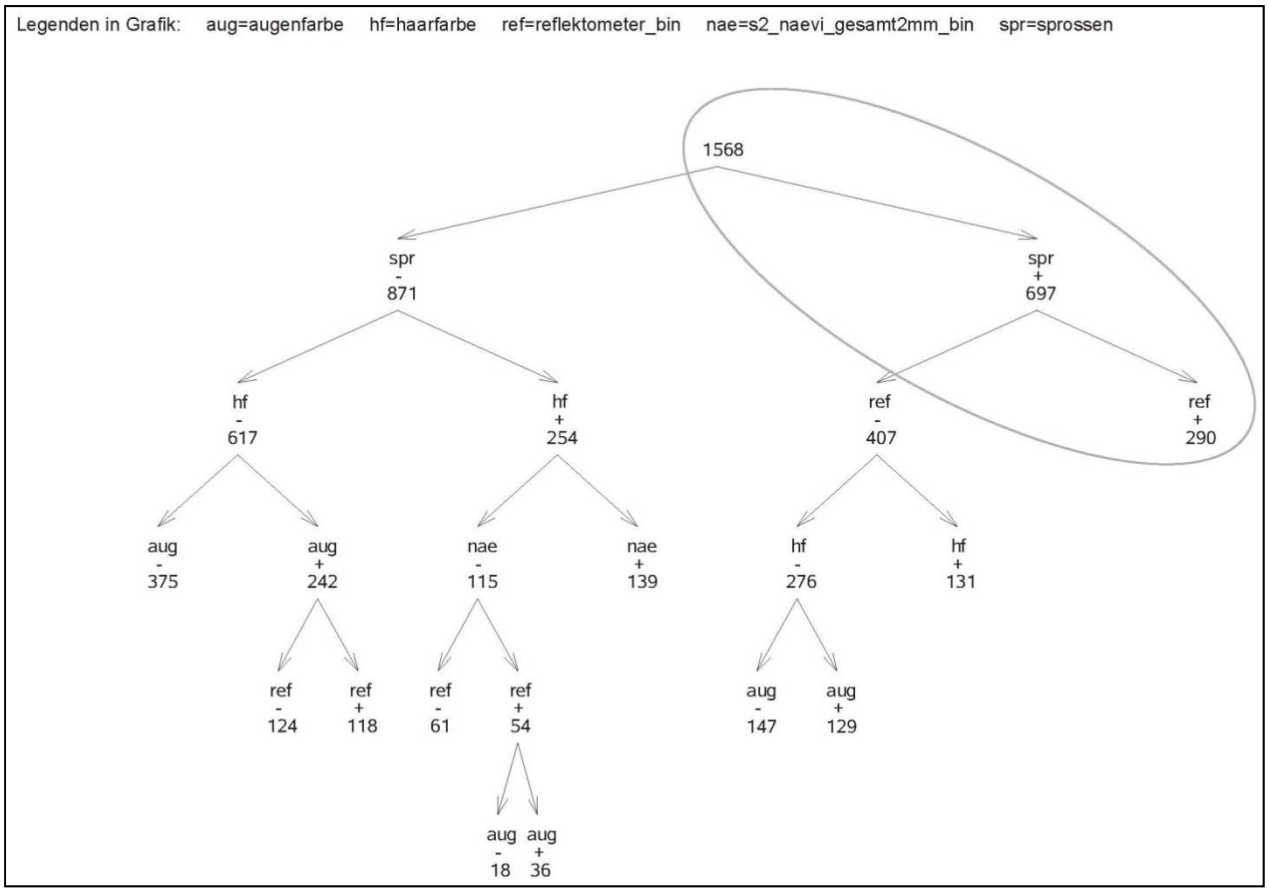

Abbildung 5: ODS Grafik über den gefundenen binären Klassifikationsbaum mit Markierung der für uns relevanten Ergebnisse

Die eigentliche Klassifikation findet immer in den Endknoten (Knoten ohne Nachfolger) statt. Es sollen als Beispiel einer Ergebnis-Interpretation der Baumgrafik nur die für uns wichtigsten Resultate der Analyse für den rechten positiven (Kennzeichnung in Grafik mit +) Hauptast des Klassifikationsbaums betrachtet werden (s. Abb. 5, Markierung). An oberster Stelle steht der sogenannte Wurzelknoten (Basis) mit den gesamten N's (in unserem Datensatz: N=1568 Beobachtungen). Die exponierte Gruppe, die wir in der Analyse betrachten, ist die Gruppe des UV-lichtempfindlicheren Hauttyps (`helligkeit=1`). Die Klassifikation für diese Gruppe ist gesucht. Somit ergibt sich für den ersten Split für Sommersprossen (spr +): auf der ersten Baumebene rechts (s. Abb. 5) Anzahl Getesteter n=697 mit Sommersprossen (`sprossen=1`, vgl. dazu auch Abb. 4, Tab. B, zweite Zeile). Was bis zu dieser Stelle wieder bedeutet, dass Sommersprossen wichtigstes Kriterium für die Klassifikation in die Gruppe mit hoher UV-Lichtempfindlichkeit ist. Ab der zweiten Baumebene erfolgt die kombinierte Betrachtung mit dem Reflektometer-Messwert (s. Abb. 5, *ref + 290* in dritter Textzeile des Baumes): das zweite Kriterium für die Klassifikation in die Gruppe mit hoher Hautempfindlichkeit sind höhere Reflektometer-Messwerte (`reflektometer_bin=1` für insgesamt n=290, vgl. dazu auch Abb. 4, Tab. B, sechste Zeile). Zusammengefasst sind nun als Kriterien für die Klassifikation in die Gruppe mit einer hohen UV-Lichtempfindlichkeit als erstes

Sommersprossen und als zweites ein höherer Reflektometer-Messwert (Messwert > 40) ausschlaggebend.
Verknüpft mit dem Ergebnisauszug aus Abb. 4, können die Befragten mit Sommersprossen und höherem Reflektometer-Messwert mit einer Wahrscheinlichkeit von 61% in die Gruppe der UV-lichtempfindlicheren klassifiziert werden (s. Abb. 4, unterste Zeile „Anzahl getesteter Fälle“ = 0.61). Dieser Anteil von 0.61 zur Klassifikation in die als exponiert betrachtete Gruppe ist übrigens auch der höchste innerhalb unseres gesamten Klassifikationsbaums.

5 Diskussion

Durch die p-Wert basierte CART-Analyse mit dem SAS Makro „chiklass“ konnte eine optimale Klassifikation für die für uns exponierte Gruppe der UV-lichtempfindlicheren Personen (Fitzpatrick-Hauttyp I + II) gefunden werden.
Als ein für uns interessantes und zugleich wichtiges Ergebnis fanden wir durch die Analyse heraus, dass nicht wie vermutet ein höherer Reflektometer-Messwert in erster Linie zur Klassifikation in die UV-lichtempfindlicheren Hauttypen beiträgt, sondern das Vorhandensein von Sommersprossen. Unsere ursprüngliche These lässt sich so begründen, dass die Reflektometer-Messung für die Bestimmung der Hauthelligkeit eine bewährte und verlässliche Methode darstellt: Je höher der Messwert, desto heller und empfindlicher ist die Haut.

Um noch detailliertere Ergebnisse zu erhalten, können alle kategoriellen Merkmale mit mehr als zwei Ausprägungen (wie z. B. die Haarfarbe: rote, blonde, braune und schwarze Haare) durch ein SAS Makro für ordinale Merkmale (Prädiktoren) weiter differenziert werden.

6 Ausblick

Die ODS Grafik wird noch erweitert, um eine bessere Interpretation zu erhalten. Dazu sollen zusätzlich pro Knoten zu den n’s die „Anteile der getesteten Fälle“ integriert werden. Für eine noch bessere Darstellung können die Anteile der UV-lichtempfindlicheren Personen in % zur Klassifikation pro Knoten auch an eine x-Achse angepasst werden (z. B. mit einer Skala von 0% - 100%).
Des Weiteren soll das Makro dahingehend erweitert werden, dass zusätzlich zu den dichotomen Variablen auch kategorielle Merkmale in die Analyse eingehen. Die Kategorien werden dann alle innerhalb der Analyse berücksichtigt. So kann für jede der Ausprägungen eines Merkmals ein Split gefunden werden, um am Ende der Analyse eine exaktere Klassifikation zur erhalten, wie z. B. für unser vierstufiges Merkmal Haarfarbe.

Literatur

[1] T. B. Fitzpatrick: The validity and practicality of sun-reactive skin types I through VI. Arch Dermatol, 124: 869-871, 1988.

[2] W. Uter, A. Pfahlberg, B. Kalina, K. F. Kölmel, O. Gefeller: Inter-relation between variables determining constitutional UV sensitivity in Caucasian children. Photodermatol Photoimmunol Photomed, 20: 9-13, 2004.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone: Classification and Regression Trees. Wadsworth Inc., Monterey, California – USA, 1984.

[4] J. Hartung, B. Elpelt, K. H. Klösener: Statistik, Lehr- und Handbuch der angewandten Statistik. R. Oldenbourg Verlag, München, Wien, 12. Auflage, 1999.

[5] I. Guggenmoos-Holzmann, K. D. Wernecke: Medizinische Statistik. Blackwell Wissenschaftsverlag, Berlin-Wien, Oxford, Edinburgh, Boston, Melbourne, Paris, Yokohama, 1996.

[6] SAS Institute Inc.: SAS Macro Language Reference, Version 8. SAS Institute Inc, Cary, NC – USA, 1999.