Ein Algorithmus zur Auswahl einer vollständigen Datenmenge

Bernd Paul Jäger
Ernst-Moritz-Arndt-Universität Greifswald, Institut für Biometrie und Med. Informatik
Walther-Rathenau-Straße 48
Greifswald
bjaeger@biometrie.uni-greifswald.de

Michael Wodny
Ernst-Moritz-Arndt-Universität Greifswald, Institut für Biometrie und Med. Informatik
Walther-Rathenau-Straße 48
Greifswald
wodny@biometrie.uni-greifswald.de

Sandra. Lieckfeldt
Ernst-Moritz-Arndt-Universität Greifswald, Institut für Biometrie und Med. Informatik
Walther-Rathenau-Straße 48
Greifswald
sandera24@web.de

Philipp Otto
Ernst-Moritz-Arndt-Universität Greifswald, Institut für Biometrie und Med. Informatik
Walther-Rathenau-Straße 48
Greifswald
ottoph@uni-greifswald.de

Paul Eberhard Rudolph
Leibniz-Institut für Nutztierbiologie, FB Genetik und Biometrie
Wilhelm-Stahl-Allee 2
Dummerstorf
pe.rudolph@fbn-dummerstorf.de

Karl-Ernst Biebler
Ernst-Moritz-Arndt-Universität Greifswald, Institut für Biometrie und Med. Informatik
Walther-Rathenau-Straße 48
Greifswald
biebler@biometrie.uni-greifswald.de

Zusammenfassung

Vorgestellt wird ein SAS-Programm, das in der Lage ist, aus einer mit Ausfallwerten behafteten großen Datei alle vollständigen Dateien auszuwählen, wie sie für zahlreiche statistische Verfahren benötigt werden. Die letzte Entscheidung darüber, für welche der vollständigen Dateien man sich entscheidet und welche Variablen man unbedingt in der Analyse haben möchte, liegt aber in der Hand des Nutzers.
Diese letzte Entscheidung sollte nicht automatisiert werden, etwa in der Art, dass das Produkt aus Variablenanzahl und nutzbaren Datensätzen maximal wird. Im vorgestellten Beispiel stellte sich im Nachhinein heraus, dass dieses Optimum zum Einsatz kam.
Bei der Diskriminanzanalyse mit ab- oder aufbauenden Verfahren der Variablenselektion ergab sich als Nebenprodukt eine Arbeitsstichprobe, die nicht in der Lernstichprobe enthalten war und das Verfahren durch „echte“ Klassifikation zu testen gestattete.

Schlüsselwörter: vollständige Datenmenge, Variablenselektion, Diskriminanzanalyse, PROC IML

1 Einleitung

Viele Prozeduren der Statistik benötigen eine vollständige Datenmenge, d. h. für alle Datensätze darf keine Variable fehlende Werte enthalten. Große Datenmengen enthalten naturgemäß auch zahlreiche Missings. Deshalb ist die obige Forderung eine einschränkende, zumindest lästige Forderung. Zahlreiche Varianten sind ersonnen worden, um fehlende Werte zu ergänzen, etwa durch Einsetzen von Gruppenmitteln oder durch eine normalverteilte Zufallszahl mit dem Gruppenmittel und der Gruppenvarianz der vorhandenen Messwerte. Diese Art der „Datenfälschung“ soll hier nicht behandelt werden.
Nimmt man einzelne Datensätze oder Variablen aus der Analyse heraus, kann man in der Regel Vollständigkeit erreichen. Es wird hier eine Vorgehensweise aufgezeigt, die das „Probieren“ des Streichens besonders lückenhafter Datensätze und Variablen ersetzt durch eine objektive Suche nach einer geeigneten vollständigen Datenmenge. Das Verfahren wird als SAS-Programm vorgestellt und an einem Beispiel erläutert.
Wendet man auf diese vollständige Datenmenge (Lernstichprobe) eine Diskriminanzanalyse mit Variablenselektion an, werden in der Regel nicht alle ins Verfahren eingespeisten Variablen benötigt. Die resultierende eingeschränkte Variablenmenge definiert eine vollständige Datenmenge größeren Umfangs. Diese Datenmenge zerfällt in natürlicher Weise in Lern- und Arbeitsstichprobe. Damit können das Klassifikationsergebnis überprüft und die tatsächlichen Fehler geschätzt werden ohne Bootstrap-Methoden zu bemühen.

2 Die medizinische Aufgabenstellung

Es wurden 21 Labordaten des Universitätsklinikums Greifswald vom Jahr 2005 verwendet, die eine Bedeutung für die Nierendiagnostik haben. Diese wurden aus den standardmäßig bei Einlieferung ins Klinikum erhobenen Labordaten ausgewählt.
Zwei Patientengruppen sind in die Analyse eingegangen, die Gruppe der chronisch Nierenerkranken CNK und als Vergleichsgruppe G diejenigen Patienten, wo sich weder in der Hauptdiagnose noch in einer der 49 Nebendiagnosen ein Hinweis auf Nierenerkrankungen ergab. Eine weitere Gruppe stellen die Patienten dar, die keine chronische Nierenerkrankung haben, aber in Haupt- oder Nebendiagnosen eine andere Nierenerkrankung oder eine Krankheit mit Nierenbeteiligung haben. Die Diagnosen in dieser Gruppe sind zu heterogen, um eine eigene Gruppe „nierenkrank, aber nicht chronisch nierenkrank“ zu definieren. Diese Patienten wurden nicht in die Analyse aufgenommen.
In der Regel werden Laborwerte mehrfach bei einem Krankenhausaufenthalt kontrolliert und manche Patienten kamen 2005 darüber hinaus auch mehrfach in das Universitätsklinikum, unter Umständen mit neuen Erkrankungen. Dann wurde der jeweils erste Krankenhausaufenthalt ausgewählt unter der Annahme, dass die Werte beim ersten Aufenthalt am wenigsten durch die Behandlung verändert sind. Übrig blieb pro Patient ein einziger Parametersatz. Insgesamt sind es 19464 Datensätze. Möglicherweise gehört der Datensatz zu einem solchen Patienten, bei dem erst bei einem späteren Krankenhausaufenthalt im Jahr die Diagnose „Chronische Nierenerkrankung „Stadium 1“ ge-

stellt wurde und die Laborparameter den Zustand vor Krankheitsbeginn beschreiben. Diese aufwändige Vorselektion wird hier nicht besprochen.
Mit einer Diskriminanzanalyse soll dann versucht werden, einen Patienten allein unter Berücksichtigung der 21 Laborparameter zu klassifizieren. Wenn das erfolgreich wäre und ohne größere Klassifikationsfehler gelänge, wäre es möglich, unmittelbar bei der Einweisung eines Patienten vom Labor aus den Hinweis auf eine chronische Nierenerkrankung CNK zu geben.

3 Beschreibung des Algorithmus zur Auswahl einer vollständigen Datenmenge

Leider sind die Datensätze nicht vollständig, stets fallen einige Laborwerte aus. Die Diskriminanzanalyse erfordert aber eine vollständige Datenmenge. Zahlreiche Varianten sind ersonnen worden, um fehlende Werte zu ergänzen:

- Das Ersetzen eines fehlenden Wertes durch das Gruppenmittel führt dazu, dass zwar der Mittelwert in den Gruppen unverändert bleibt, die Varianz der entsprechenden Zufallsvariable jedoch mit zunehmender Anzahl an Ersetzungen monoton verkleinert wird.
- Ersetzt man die Missingwerte durch eine normalverteilte Zufallszahl mit dem Gruppenmittel und der Gruppenvarianz der vorhandenen Messwerte, so ignoriert man die Kovarianz mit anderen Variablen.

Diese Art der „Datenfälschung“ soll hier nicht ausgeführt werden.

Die Datei mit 19464 Datensätzen überblicken zu wollen, ist ebenso aussichtslos, wie durch Erraten auf eine möglichst große Datenmenge zu kommen, bei der auf der einen Seite möglichst viele der 21 Variablen, auf der anderen aber auch möglichst viele der Patienten eingehen sollen.
Vorausgesetzt wird für den Algorithmus eine SAS-Datei X, die ausschließlich aus numerischen Variablen besteht. Das ist für die 21 Laborwerte der Fall. Sollten alphanumerische Variablen vorhanden sein, müssen diese transformiert werden. Beispielsweise wird die Variable GESCHLECHT mit den Ausprägungen „männlich“ und „weiblich“ zur Variable GESCHLECHT1 umgewandelt, in der eine 1 steht, wenn GESCHLECHT = „männlich“ und eine 0 steht, wenn GESCHLECHT =„weiblich“.
Hat die alphanumerische Variable mehr als zwei Kategorien, sollte man diese nicht durch ganze Zahlen kodieren, sondern im Hinblick auf eine spätere Weiterbearbeitung mit der Diskriminanzanalyse so genannte Dummy-Variablen einführen. Hat beispielsweise die Variable WUNDAUSDEHNUNG die Kategorien „klein“, „mittel“ oder „groß“, gewinnt man daraus drei neue Variablen WUNDE_KLEIN, WUNDE_MITTEL und WUNDE_GROSS, die jeweils mit einer 1 ausgefüllt werden, wenn die Variable Wundausdehnung die entsprechende Kategorie aufweist. Ansonsten wird sie mit einer 0 oder einem fehlenden Wert versehen, je nachdem ob eine andere Kategorie vorliegt bzw. die Wundausdehnung für diesen Datensatz nicht erhoben wurde.

3.1 Programmschritt 1

Die Datei g.laborwerte mit den Laborwerten wird mittels PROC IML in einer Matrix x erfasst. Für jeden Datensatz i ($1 \leq i \leq z$) und jede Variable j ($1 \leq j \leq s$) wird der Wert in ein Dualwort W[i,j] umgewandelt. Das Dualwort ist eine 1, wenn der Datensatz i bezüglich der j-ten eingelesenen Variable vorliegt und eine 0, wenn es ein fehlender Messwert ist. Eine SAS-Datei `work.asdf`, bestehend aus diesen Dualworten aller Datensätze, wird ausgegeben.

Programmschritt 1:

```
libname g "G:\Niere ";
%let s=21;
proc iml;
use g.laborwerte;
read all var _num_ into x;
*liest alle numerischen Variablen in ein Feld x;
z=nrow(x); *nrow = Anzahl der Zeilen;
W=J(z,&s,"A");
do i=1 to z;
do j=1 to &s;
if x[i,j]^=. then W[i,j]="1"; else W[i,j]="0";
*x wird in ein alphanumerisches Feld der Länge s (Anzahl der
Variablen) umgewandelt, bestehend aus den Zahlen 1 und 0;
end;
end;
create asdf from W;
append from W; *Ausgabe in eine SAS-Datei work.asdf;
quit;
run;
```

3.2 Programmschritt 2

In einem Datastep werden die s Dualworte, die zu einem Datensatz gehören, zu einem Dualwort WORD der Länge s = 21 verkettet und mit der PROC FREQ wird ausgezählt, wie häufig jede Variante von WORD vorkommt. Die Ausgabe der PROC FREQ wird angezeigt, ist lexikografisch geordnet und wird gleichzeitig in eine SAS-Datei WORK.HELP umgeleitet.

Programmschritt 2:

```
data asdf;
set asdf;
Wort=cat(Of col1-col&s);*Verketten des Datensatzes zu einem Wort;
run;
proc freq data=asdf;
tables wort/out=help ;
run;
data help;
set help;
keep wort count;
run;
```

Tabelle 1: Auszählung der Häufigkeiten der Worte aus der PROC FREQ im Programmschritt 2 (Ausschnitt aus der Gesamttabelle)

Beob.	Wort	Anzahl	Prozent
1	00000000000000000100	1	0.0051
2	00000000000001000000	2	0.0103
3	00000000100000000000	9	0.0462
4	00000000100001000000	9	0.0462
5	00000000110001000000	3	0.0154
6	**00000001000000000010**	**15**	**0.0771**
7	**00000001000000000110**	**25**	**0.1284**
8	**00000001010000000110**	**1**	**0.0051**
9	**00000001100001000010**	**1**	**0.0051**
….	…..	….	…
45	00011101111111000011	24	0.1233
46	00011101111111000101	62	0.3185
47	**00011101111111000111**	**1057**	**5.4305**
48	00011101111111001111	1	0.0051
49	00011101111111100111	3	0.0154
50	00011101111111101011	1	0.0051
….	….	….	….
117	**01011101111111000111**	**1958**	**10.0596**
118	01011101111111001011	1	0.0051
119	01011101111111001011	5	0.0257
120	01011101111111010011	1	0.0051
….	….	….	….
151	**01011111111111000111**	**3004**	**15.4336**
152	01011111111111001011	2	0.0103
153	01011111111111001011	20	0.1028
….	….	….	….
299	11011111111111000011	44	0.2261
300	11011111111111000101	213	1.0943
301	**11011111111111000111**	**4154**	**21.3420**
302	11011111111111001011	1	0.0051
303	11011111111111001011	4	0.0206
304	11011111111111001011	39	0.2004
….	….	….	….
334	11111111111111100111	82	0.4213
335	11111111111111101011	1	0.0051
336	11111111111111101011	1	0.0051
337	11111111111111101011	14	0.0719
338	11111111111111101111	2	0.0103
339	11111111111111100111	2	0.0103

Nach dem ersten und zweiten Programmschritt stellt man fest, dass nur ein kleiner Anteil der denkbaren 2^{21} = 2.097.152 Möglichkeiten realisiert wurde, nämlich 339. Selbstverständlich sind die Worte, die mit einer hohen Anzahl an Datensätzen einhergehen, verdächtig, dass ihre Variablenkombination auch zu einer optimalen Dateigröße führt. Leider ist das nur ein notwendiges Kriterium, kein hinreichendes.

3.3 Programmschritt 3

Wenn man das 6. bzw. 7. Wort mit den Häufigkeiten 15 bzw. 25 der Tabelle 1 ansieht, ist erkennbar, worauf sich das hinreichende Kriterium beziehen muss.
Das 6. Wort hat an der 8. und 20. Position eine 1. Das bedeutet, dass es genau 15 Datensätze gibt, bei denen ausschließlich die 8. und 20. Variable ausgefüllt sind. Das 7. Wort hat ebenfalls an der 8. und 20. Position eine 1 und darüber hinaus auch noch an der 19. Bei optimalen Datenmengen interessiert man sich für solche Datensätze die *mindestens an den vorgegebenen Positionen*, nicht an *genau den Positionen*, einen Eintrag besitzen. Es gibt demnach mindestens 40 Datensätze, bei denen an der Position 8 und 20 eine 1 steht. Bei aufmerksamen Durchmustern findet man weitere Worte, die an 8. und 20. Position eine 1 stehen haben. Dazu zählen die beiden unmittelbar folgenden und möglicherweise noch viele weitere der 339 anderen Worte.
Um alle herauszufinden, wird die in der PROC FREQ im Programmschritt 2 ausgegebene Datei work.help in PROC IML eingelesen.
Man bildet nacheinander für jedes Wort, das als ein Vektor der Länge s = 21 aufgefasst wird, das Skalarprodukt $< . ; . >$ mit sich selbst. Beispielsweise

<W6;W6> = < (0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0);

(0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0) > = 2.

Alle weiteren Worte Wi mit <W6,Wi> = 2 haben mindestens an den Positionen, an denen W6 eine 1 stehen hatte ebenfalls eine 1. Das gilt für die oben erwähnten Worte W7, W8 und W9:

<W6;W7> = < (0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0);
(0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,1,0) > = 2,
<W6;W8> = < (0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0);
(0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0,1,0) > = 2 und
<W6;W9> = < (0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0);
(0,0,0,0,0,0,0,1,1,0,0,0,0,1,0,0,0,0,0,1,0) > = 2 .

Die Häufigkeiten der weiteren Worte Wi mit <W6; Wi> = 2 müssen zur Häufigkeit, mit der W6 auftritt, hinzugezählt werden. Da die Worte in der Datei work.help aus der PROC FREQ stammen, sind sie bezüglich der lexikografischen Ordnung aufsteigend geordnet und man muss nur Worte Wi berücksichtigen, die in der Datei unterhalb liegen, im Beispiel i > 6 (siehe Programmzeile DO i=j+1 to n;).

Programmschritt 3:

```
Proc IML;
use help;
read all var {wort} into x;
read all var {count} into count;
 n =NROW(x); *Anzahl der Sätze;
 sl=LENGTH(x[1,1]); *Länge der Strings;
 M=J(n,sl,.);
 vanz=J(n,1,0);
 DO j=1 to n;
  DO i=1 to sl;
   if SUBSTR(x[j,1],i,1)="0" then M[j,i]=0;
   ELSE M[j,i]=1;
   vanz[j]=(M[j,+]); *Anzahl der belegten Zellen pro Zeile;
  END;
 END;
 DO j=1 to n;
  Sp=(M[j,]#M[j,])[+]; *Skalarprodukt mit sich selbst;
  DO i=j+1 to n;
   *Skalarprodukt mit anderen Zeilen --> Erhöhung der Häufig-
keit;
   if Sp=(M[j,]#M[i,])[+] Then count[j]=count[j]+count[i];
  END;
 END;
create help2 from count;
append from count;
create help3 from M;
append from M;
create help4 from vanz;
append from vanz;
quit;
run;
```

3.4 Programmschritt 4

In einem Datastep werden die von der PROC IML ausgegebenen SAS-Dateien mit MERGE zur Datei work.werte_anzahl vereinigt. Die Variable Produkt, die durch Multiplikation der Variablenanzahl und den entsprechenden Datensätzen entsteht, ist ein Maß für die optimale Größe der verbleibenden Datei. Mit der PROC MEANS wird das Maximum bestimmt und ausgegeben. Für mindesten 13 Variablen liegen bei 13523 Datensätzen Messwerte vor, das sind 175799 Dateneintragungen. Das zugehörige Wort

(0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1)

entspricht den ausgefüllten Variablen

V4, V5, V6, V8, V9, V10, V11, V12, V13, V14, V19, V20, V21.

Neben diesem Maximum wird eine Gesamtübersicht gegeben, geordnet nach Variablenanzahl und Datensätzen, damit der Nutzer unter Umständen eine andere Variante mit mehr Variablen oder mehr Datensätzen auswählen kann. Für die durchgeführte Diskriminanzanalyse wurden diese 13 Variablen ausgewählt.

Tabelle 2: Gesamtübersicht

Var	Datensätze	Prod.	Variablennummer																				
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	348	348	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2 930	2 930	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1	9 043	9 043	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	12 556	12 556	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	15 176	15 176	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	15 735	15 735	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
1	17 048	17 048	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
1	19 196	19 196	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	8 161	16 322	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	12 469	24 938	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
7	19 192	134 344	0	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1
8	15 031	120 248	0	1	0	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1
8	15 580	124 640	0	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1	0	1
8	15 777	126 216	0	0	0	1	1	1	0	1	0	0	1	1	1	0	0	0	0	0	0	0	1
8	16 859	134 872	0	0	0	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	0	0	1
8	17 294	138 352	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	1
9	11 295	101 655	0	1	0	1	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0	0	1
9	12 335	111 015	0	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	1
9	12 820	115 380	0	1	0	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1	0	1
9	13 871	124 839	0	0	0	1	1	1	0	1	0	1	1	1	1	0	0	0	0	0	0	1	0
9	14 340	129 060	0	1	0	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	0	0	1
9	14 759	132 831	0	0	0	1	1	1	0	1	0	0	1	1	1	0	0	0	0	0	0	1	1
9	15 536	139 824	0	0	0	1	1	1	0	1	0	0	1	1	1	0	0	0	0	0	1	0	1
9	17 211	154 899	0	0	0	1	1	1	0	0	1	0	1	1	1	1	0	0	0	0	0	0	1
10	10 544	105 440	0	0	0	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	1	0	1
10	11 231	112 310	0	1	0	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	1
10	12 165	121 650	0	0	0	1	1	1	1	0	0	1	1	1	1	1	0	0	0	0	0	0	1
10	12 786	127 860	0	1	0	1	1	1	0	1	0	0	1	1	1	0	0	0	0	0	1	0	1
10	13 869	138 690	0	0	0	1	1	1	0	1	0	1	1	1	1	0	0	0	0	0	0	1	1
10	14 166	141 660	0	1	0	1	1	1	0	0	0	1	1	1	1	1	0	0	0	0	0	0	1
10	14 452	144 520	0	1	0	1	1	1	0	0	1	0	1	1	1	1	0	0	0	0	0	0	1
10	14 533	145 330	0	0	0	1	1	1	0	1	0	0	1	1	1	0	0	0	0	0	1	1	1
10	14 565	145 650	0	0	0	1	1	1	0	1	0	1	1	1	1	0	0	0	0	0	1	0	1
10	15 159	151 590	0	0	0	1	1	1	0	1	1	0	1	1	1	1	0	0	0	0	0	0	1
10	16 513	165 130	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	1
11	10 031	110 341	0	0	0	1	1	1	1	1	0	0	1	1	1	0	0	0	0	0	1	1	1
11	11 072	121 792	0	1	0	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1
11	11 112	122 232	0	1	0	1	1	1	1	0	1	0	1	1	1	1	0	0	0	0	0	0	1
11	11 117	122 287	0	1	0	1	1	1	1	0	0	1	1	1	1	1	0	0	0	0	0	0	1
11	11 932	131 252	0	1	0	1	1	1	0	1	0	0	1	1	1	0	0	0	0	0	1	1	1
11	12 112	133 232	0	0	0	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0	0	0	1
11	12 452	136 972	0	1	0	1	1	1	0	1	0	1	1	1	1	0	0	0	0	0	1	0	1
11	12 647	139 117	0	1	0	1	1	1	0	0	1	0	1	1	1	1	0	0	0	0	1	0	1
11	13 674	150 414	0	0	0	1	1	1	0	1	0	1	1	1	1	0	0	0	0	0	1	1	1
11	14 111	155 221	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	1
11	14 200	156 200	0	0	0	1	1	1	0	1	1	0	1	1	1	1	0	0	0	0	0	1	1
11	14 441	158 851	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0	0	0	0	1	0	1
11	14 608	160 688	0	0	0	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	0	0	1
11	14 940	164 340	0	0	0	1	1	1	0	1	1	0	1	1	1	1	0	0	0	0	1	0	1
12	9 211	110 532	0	1	0	1	1	1	1	1	0	0	1	1	1	0	0	0	0	0	1	1	1
12	9 274	111 288	0	1	0	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	1	1
12	9 686	116 232	0	1	0	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	1	0	1
12	9 989	119 868	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	1	1	1
12	11 070	132 840	0	1	0	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0	0	0	1

12	11 661	139 932	0	1	0	1	1	1	0	1	0	1	1	1	1	0	0	0	0	0	1	1
12	11 754	141 048	0	1	0	1	1	1	0	1	0	1	1	1	1	1	0	0	0	0	1	1
12	11 942	143 304	0	1	0	1	1	1	0	1	1	0	1	1	1	1	0	0	0	0	1	1
12	12 392	148 704	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0	0	0	0	1	0	1
12	12 514	150 168	0	1	0	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	0	0	1
12	12 614	151 368	0	1	0	1	1	1	0	1	1	0	1	1	1	1	0	0	0	0	1	0	1
12	13 525	162 300	0	0	0	1	1	1	0	1	1	1	1	1	1	0	0	0	0	0	1	1	1
12	13 561	162 732	0	0	0	1	1	1	0	1	0	1	1	1	1	1	0	0	0	0	1	1	1
12	13 713	164 556	0	0	0	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	0	1	1
12	13 995	167 940	0	0	0	1	1	1	0	1	1	0	1	1	1	1	0	0	0	0	1	1	1
12	14 405	172 860	0	0	0	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	1	0	1
13	9 181	119 353	0	1	0	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	1	1	1
13	9 233	120 029	0	1	0	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	1	1
13	9 234	120 042	0	1	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	1	1
13	9 637	125 281	0	1	0	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0	1	0	1
13	9 644	125 372	0	1	0	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	1
13	9 709	126 217	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1
13	9 926	129 038	0	0	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0	0	1	1	1
13	9 928	129 064	0	0	0	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	1	1
13	10 005	130 065	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1
13	10 422	135 486	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	1
13	11 608	150 904	0	1	0	1	1	1	0	1	0	1	1	1	1	1	0	0	0	0	1	1	1
13	11 720	152 360	0	1	0	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	0	1	1
13	11 792	153 296	0	1	0	1	1	1	0	1	1	0	1	1	1	1	0	0	0	0	1	1	1
13	12 362	160 706	0	1	0	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	1	0	1
13	**13 523**	**175 799**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**1**
14	9 142	127 988	0	1	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0	0	1	1	1
14	9 142	127 988	0	1	0	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	1	1
14	9 205	128 870	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1
14	9 615	134 610	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	1
14	9 896	138 544	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1
14	11 577	162 078	0	1	0	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	1	1	1
15	9 115	136 725	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1
19	16	304	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1
20	2	40	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1

Programmschritt 4:

```
data help2;
set help2;
keep Datensaetze; *Weglegen der Systemariablen;
datensaetze=col1; *Häufigkeit der Zeilenbelegung;
run;
data help4;
set help4;
keep anz; * Weglegen der Systemariablen;
anz=col1; * Variablenanzahl;
run;
data werte_anzahl;
merge help2 help3 help4;
produkt=anz*datensaetze; *Anzahl belegter Zellen;
run;
proc sort data=werte_anzahl;
by  anz datensaetze;
run;
title Gesamtübersicht;
proc print data=werte_anzahl noobs;
var  anz datensaetze produkt col1--col&s;
```

```
run;
title Maximum;
proc means max;
var produkt;
run;
```

4 Die Diskriminanzanalyse

Vorangestellt wurde die parametrische Prozedur PROC STEPDISC. Mit dem aufbauenden Verfahren wurden sechs Variablen ausgewählt, bis keine Verbesserung erreichbar war.
Mit diesen sechs Variablen wurde die 5-nächste-Nachbarn-Methode der PROC DISCRIM als parameterfreie Variante der Diskrimination durchgeführt. Die parametrische und parameterfreie Methode kommen in der Regel zu anderen Klassifikationsergebnissen. Trotz aller Bedenken wird die ausgewählte Parametermenge der PROC STEPDISC nahe an der optimalen Variablenmenge liegen. Durch wiederholten Austausch einer Variablen aus der ausgewählten mit einer Variable aus der nichtausgewählten Gruppe, die bei Verschlechterung des Klassifikationsergebnisses rückgängig gemacht wurde, konnte das Reklassifikationsergebnis nur unwesentlich verbessert werden. Das Endresultat enthält Tabelle 3. Erstaunlich gut werden die Patienten klassifiziert. Nicht einer der chronischen Nierenpatienten wird falsch und auch 94.4% der Vergleichspersonen werden richtig, 5.6% falsch zugeordnet.

Tabelle 3: Reklassifikationsergebnis der 5-nächste-Nachbarn-Diskriminanzanalyse mit sechs Variablen der optimaler Datenmenge (Lernstichprobe = Arbeitsstichprobe)

		in		
		G	CNK	Summe
von	G	6122	363	6485
	CNK	0	422	422
Summe		6122	785	6907

Bekanntlicherweise ist das Reklassifikationsergebnis ein wenig zu optimistisch, realistischer ist das Klassifikationsergebnis mit einer neuen Arbeitsstichprobe.
Beim Start der Diskriminanzanalyse hat man sich auf Datenvektoren bezogen, die an mindestens 13 Positionen ausgefüllt waren. Die Lernstichprobe enthielt 6907 Datensätze.
Eine Auswahl an Datenvektoren, die an genau sechs Teilpositionen der 13 eine 1 besitzen, ist weniger einschränkend und enthält dadurch mehr Datensätze, nämlich 9218. Dazu gehören selbstverständlich die obigen 6907 der Lernstichprobe, aber die 9218 – 6907 = 2311 Datensätze sind nicht in der Lernstichprobe enthalten und bilden die Arbeitsstichprobe. Das Klassifikationsergebnis dieser Arbeitsstichprobe enthält Tabelle 4.
Vergleicht man die Reklassifikation der Lernstichprobe in Tabelle 3 mit der Klassifikation der Arbeitsstichprobe, so stellt man gute Übereinstimmung fest. Die Fehlklassifikation in der Gruppe der Patienten mit chronischer Nierenerkrankung beträgt 2.5% und in der Vergleichsgruppe ist sie sogar leicht auf 4% gesunken.

Tabelle 4: Klassifikationsergebnis der 5-nächste-Nachbarn-Diskriminanzanalyse mit sechs verbleibenden Variablen in der Arbeitsstichprobe

		in		
		G	CNK	Summe
von	G	2140	89	2229
	CNK	2	80	82
Summe		2142	169	2311

5 Zusammenfassung

Vorgestellt wurde ein SAS-Programm, das in der Lage ist, alle vollständigen Teildateien aus einer hochdimensionalen großen Datei auszuwählen, wie sie für zahlreiche statistische Verfahren benötigt werden. Die letzte Entscheidung darüber, für welche der vollständigen Teildateien man sich entscheidet und welche Variablen man unbedingt in der Analyse haben möchte, liegt aber in der Hand des Nutzers.
Diese Entscheidung sollte unserer Meinung nach nicht automatisiert werden, etwa in der Art, dass das Produkt aus Variablenanzahl und nutzbaren Datensätzen maximal wird. Im vorgestellten Beispiel stellte sich im Nachhinein heraus, dass dieses Optimalitätskriterium zum Einsatz kam.
Bei der Diskriminanzanalyse mit ab- oder aufbauenden Verfahren der Variablenselektion ergab sich als Nebenprodukt eine Arbeitsstichprobe, die nicht in der Lernstichprobe enthalten war und das Verfahren durch „echte“ Klassifikation zu testen gestattet.

Literatur

[1] SAS Institute Inc., 2004. SAS/STAT 9.1 User Guide. Cary, NC: SAS Institute Inc.

[2] Wodny, M. u.a.: Statistikpraktikum mit SAS. Shaker Verlag Aachen 2010