Clusteranalyse mit Binärdaten

Bernd Jäger *, Michael Wodny *, Paul Eberhard Rudolph **,
Dana Patschinsky***

* Institut für Biometrie und Medizinische Informatik, Ernst-Moritz-Arndt-Universität Greifswald
17489 Greifswald
E-Mail: jaeger@biometrie.uni-greifswald.de

** Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere Dummerstorf/Rostock
18196 Dummerstorf
Tel. 038208-68908
E-Mail: rudolph@fbn-dummerstorf.de

***Institut für Rechtsmedizin, Ernst-Moritz-Arndt-Universität Greifswald
17489 Greifswald

1. Einleitung

Im Rahmen des Forschungsverbundes Community Medicine an der Medizinischen Fakultät der Ernst-Moritz-Arndt-Universität Greifswald werden in einem interdisziplinären Teilprojekt des Instituts für Rechtsmedizin, des Lehrstuhls für Kriminologie und des Instituts für Psychologie Fragen des Alkoholkonsums und der Straßenverkehrsdelinquenz untersucht. Die statistische Beratung und Auswertung erfolgte im Institut für Biometrie und Medizinische Informatik.
Es wurden verschiedene Laborparameter in 1015 Blutproben von auffälligen Verkehrsteilnehmern mit akutem Trunkenheitsverdacht aus dem Raum Vorpommern bestimmt, mit dem Ziel eine Gruppe von „trinkenden Fahrern“ von einer Gruppe der „fahrenden Trinker“ abzugrenzen. Möglicherweise kann man auch noch andere Einteilungen aus der Gesamtpopulation herausfinden.
Das ist der klassische Ansatz für eine Clusteranalyse, die Personen mit ähnlichem Merkmalsspektrum zu einer Gruppe zusammenfasst.
Untersucht werden neben dem Blutalkoholspiegel, der eine enge Korrelation zum Atemalkoholwert aufweist, der Methanolwert, die GGT (Gamma-Glutamyltranspeptidase) und CDT (Carbohydrate-deficient Transferrin).
Methanol ist ein Begleitstoff in vielen alkoholischen Getränken und ein Indikator für die Dauer der Alkoholisierungsphase. Ein Ethanolspiegel von über 0.2 Promille hemmt den Abbau von Methanol und führt zu einer Akkumulation von endogenem und exogen aufgenommene Methanol. Bei einem Methanolspiegel von über 10 mg/l besteht der Verdacht eines kurz- bis mittelfristigen Alkoholmissbrauchs.
Die GGT ist bei singulärer Erhöhung (ohne Erhöhung anderer leberspezifischer Laborwerte) als Ausdruck einer Enzyminduktion in der Leber typisch für einen chronischen Alkoholkonsum. Jedoch kann eine Fülle von Erkrankungen und Noxen ebenfalls zu einer Erhöhung der GGT führen. Meist wird ein Normalwert von unter 0.82 µmol/sl angegeben.
CDT ist ein abnormales Eisen transportierendes Glycoprotein, welches unter chronischem Alkoholmissbrauch mit hoher Prävalenz im Blut auftritt. Die Pathomechanismen des CDT-Anstieges sind noch nicht befriedigend aufgeklärt, es handelt sich wahrscheinlich um eine Störung im Glycoprotein/Glycolipidstoffwechsel. Eine CDT-Konzentration von über 6 % ist als spezifischer Marker für einen chronischen Alkoholkonsum anzusehen.
Die Tabelle 1 spiegelt diese unscharfe Gruppeneinteilung wider.
Neben diesen stetigen Laborparametern liegen für die untersuchten Personen zahlreiche binäre Merkmale vor, insbesondere sind das die Ergebnisse psychologischer Tests (LAST - **L**übecker **A**lkoholismus **S**creening **T**est), die noch nicht in diese Analyse eingegangen sind.

Die alternativen Merkmale Altersgruppen jung/alt (Alter <30 / Alter 30), das Geschlecht m/w, Tag/Nacht der Blutabnahme (6.00 - 18.00Uhr / 0.00 - 5.59Uhr oder 18.01 - 24.00Uhr) bzw. Wochenende ja/nein (Freitag 18.00Uhr bis Montag 6.00Uhr / Montag 6.01Uhr bis Freitag 17.59Uhr) wurden in die Analyse einbezogen. Leider gibt es für binäre ebenso für kategoriale Merkmale keine Standardmethoden zur Clusteranalyse im SAS. Es wird ein SAS-Makro vorgestellt, das stetige und binäre Merkmale gleichzeitig der Clusteranalyse zuführt.

		Methanol				Σ
		≥10mg/l		<10mg/l		
		CDT		CDT		
		≥6%	<6%	≥6%	<6%	
GGT	≥0.82 µmol/sl	136	65	69	132	402
	< 0.82 µmol/sl	57	37	157	346	384
Σ		193	102	226	478	999
		295		704		

Tab.1: Anzahl der in den Parametern Methanol, GGT und CDT auffälligen bzw. unauffälligen Personen (2x2x2-Tafel)

2. Die Untersuchungspopulation

Die Abb. 1 und 2 geben die Verteilung des Blutalkoholspiegels, gemessen in Promille, und die Altersverteilung der Untersuchungspopulation wieder. Etwa 10% der auffälligen Verkehrsteilnehmer sind Frauen, auffällig sind die Altersgruppen über 50 Jahre bei den Frauen kaum vertreten.
Von den 1015 untersuchten Personen sind 651 (64.1%) bereit über ihre Trinkgewohnheiten Auskunft zu geben (Abb. 3).

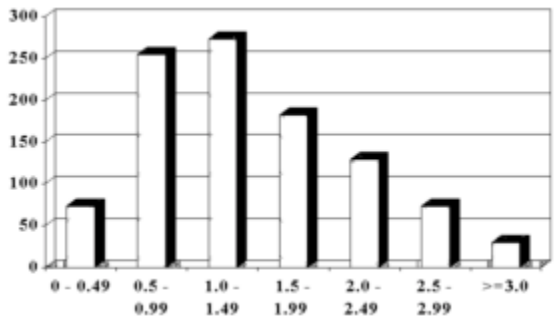

Abb. 1: Verteilung des Blutalkoholspiegels, gemessen in Promille

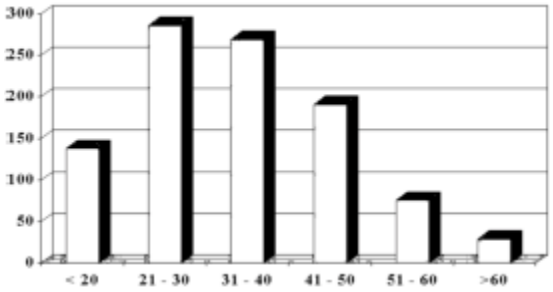

Abb.2: Altersverteilung der Untersuchungspopulation

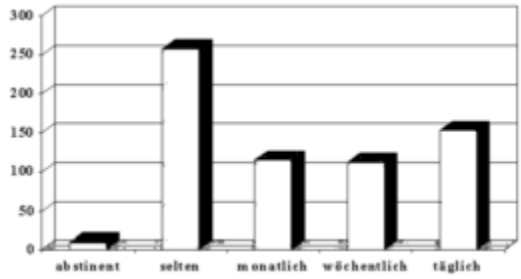

Abb. 3: Trinkgewohnheiten von 651 Auskunftswilligen

3. Die Clusteranalyse

Die Clusteranalyse ist, obwohl sie als Programm in Statistiksoftware eingebunden ist und in Lehrbüchern über mehrdimensionale statistische Verfahren beschrieben wird, kein statistisches, sondern ein algebraisch-geometrisches Verfahren, das versucht die "Ähnlichkeit" mit Hilfe von Abständen auszudrücken.

Definition Abstand oder Distanz
Es sei X eine beliebige Menge, auf der eine Funktion D(*,*) erklärt ist, die je zwei Elementen von X in eindeutiger Weise eine reelle Zahl zuordnet. Die Abbildung D heißt Abstand oder Distanz, wenn sie die folgenden Eigenschaften besitzt:

1. $D(x,y) \geq 0$ für alle $x, y \in X$ und $D(x,x)=0$ für alle $x \in X$,
2. $D(x,y) = D(y,x)$ für $x, y \in X$ und
3. $D(x,y) \leq D(x,z) + D(z,y)$ für alle $x, y, z \in X$.

Die letzte Eigenschaft wird als Dreiecksungleichung bezeichnet. Sie besagt, dass die Entfernung von je zwei Objekten nicht kleiner werden kann, wenn ein "Umweg" über einen dritten Punkt gemacht wird.
Zwei Beispiele für einen Abstand zwischen $x = (x_1, x_2, ..., x_n)$ und $y = (y_1, y_2, ..., y_n)$ aus dem ∇^n sind der Euklidische Abstand d_E ,

$$d_E(x,y)^2 = \sum_{i=1}^{n} (x_i - y_i)^2$$

oder der Abstand, der durch eine symmetrische und positiv definite Matrix A erzeugt wird

$$d_A(x,y)^2 = (x-y)\,A(x-y)^T .$$

Bestimmt man bei einer vorliegenden Untersuchungsdatei die empirische Kovarianzmatrix *COV*, und wählt als Matrix A die inverse Kovarianzmatrix COV^{-1}, so erhält man den sogenannten Mahalanobis-Abstand d_M.
Damit geht die Abhängigkeitsstruktur der Variablen der Untersuchungsdatei in die Abstandsdefinition ein. Sind die eingehenden Variablen unabhängig, so geht der Mahalanobis-Abstand in den Euklidischen Abstand über. Die inverse Kovarianzmatrix ist in diesem Falle die Einheitsmatrix. Im vorgestellten Programm wurde der Mahalanobis-Abstand dem Euklidischen vorgezogen.
Für binäre p-Tupel - mit anderen Worten endliche 0-1-Folgen der Länge p - wäre beispielsweise die sogenannte "simple matching distance" ein möglicher Abstand. Die Definition soll an einem Beispiel demonstriert werden:

Für x = (0, 0, 0, 1, 1, 0, 1, 0) ε $\{0,1\}^8$ und
y = (0, 1, 0, 1, 1, 0, 0, 0) ε $\{0,1\}^8$

gibt die Tabelle 2 an, dass zweimal (α) die Koordinatenausprägungen 1 von y mit der 1 von x übereinstimmen, nämlich an der 4. Und 5. Koordinate. Analog bestimmt β=1 die Anzahl der Ausprägung 1 von x mit der 0 von y, usw.

		Y		
		1	0	Σ
x	1	α = 2	β = 1	3
	0	γ = 1	δ = 4	5
	Σ	3	5	p = 8

Tab. 2: Gegenüberstellung der Ausprägungen der acht Koordinaten von x und y

Als Abstand wird definiert

$$d_{SM}(x,y)^2 = 1 - \frac{\alpha + \delta}{p}.$$

Damit gilt für obige x und y $d_{SM}(x,y)^2 = 1 - \frac{6}{8} = \frac{1}{4}$.

Falls eine vollständige Übereinstimmung vorliegt ist $\alpha + \beta = p$ und somit $d_{SM}(x,y) = 0$.
Darüber hinaus gibt es noch zahlreiche andere Abstände und abstandsähnliche Begriffe , die auf obiger Vierfeldertafel definiert sind, z.B. den Jaccard-Abstand (auch Tanimoto-Abstand),

$$d_J(x,y) = 1 - \frac{\alpha}{\alpha + \beta + \gamma} = 1 - \frac{2}{4} = \frac{1}{2}$$

der im vorgestellten Programm verwendet wurde, und den Czekanowski-"Abstand" ,

$$d_C(x,y) = 1 - \frac{2\alpha}{2\alpha + \beta + \gamma} = 1 - \frac{4}{6} = \frac{1}{3}$$

der allerdings nicht der Dreiecksungleichung genügt, wie man sich leicht durch ein Gegenbeispiel überzeugen kann. Trotz dieses Mangels sind solche "Abstandsfunktionen" oder Pseudoabstände in Gebrauch, weil sie eine stärkere Wichtung einer Kategorie vornehmen, die manchen Untersuchern nötig erscheint.
Einen Überblick über weitere Abstandsdefinitionen findet man bei SPÄTH(1977). In seiner Arbeit sind 14 verschiedene Ähnlichkeitskoeffizienten angegeben, die zu Abstandsdefinitionen verwandt werden.

Der Mahalanobis-Abstand variiert auf der positiven Achse. Der Jaccard-Abstand bildet in das Intervall [0, 1] ab. Damit eine gemeinsame Verwendung beider Abstände durch eine additive Verknüpfung zu einem Gesamtabstand sinnvoll wird, muss eine Normierung des Mahalanobis-Abstandes vorgenommen werden. Der Gesamtabstand d_G ist danach

$$d_G(x,y) = \frac{d_M(x,y)}{1+d_M(x,y)} + d_J(x,y)\,.$$

Die Bezeichnung x und y für die aus stetigen und binären Daten gemischten Datentupel auf der linken Seite der Gleichung ist dabei auf die rechte Seite übertragen worden, obwohl beim Mahalanobisabstand d_M nur die stetigen Koordinaten von x und beim Jaccard-Abstand d_J nur die binären Koordinaten gemeint sind.
Der Gesamtabstand variiert im Intervall [0, 2). Die Normierung des Mahalanobis-Abstandes erhält die Abstandseigenschaften, ebenso wie die Summanden diese auf die Summe vererben.

Mit der SAS-Prozedur CLUSTER sind 11 verschiedene Clustermethoden abarbeitbar, die jeweils durch spezielle Parametersätze weiter differenziert werden können. Alle diese Methoden basieren auf hierarchischen Verfahren. Anfangs ist jede Beobachtung ein Cluster. Die zwei dichtesten Cluster werden verschmolzen. Die Prozedur kann so oft wiederholt werden, bis nur ein Cluster übrig bleibt. In der SAS-Prozedur TREE wird eine Ausgabedatei der Prozedur CLUSTER genutzt, um mit einer Baumstruktur das Clusterverfahren zu illustrieren. Bei der großen Anzahl an Datensätzen stößt dieses Programm allerdings an seine Grenzen. Die Prozedur ist aber der Vollständigkeit wegen bei der Programmierung in das Makro eingearbeitet worden.

4. Das SAS - Programm zur Verarbeitung von binären und stetigen Merkmalen

Voraussetzung für das Abarbeiten des Makro ist neben den SAS-Standardmodulen die Prozedur IML. Prinzipiell kann man aber auch mit zahlreichen Data-Steps die gleichen Schritte abarbeiten, allerdings nicht so allgemeingültig, automatisch und ohne Eingriffe.
Die Aufteilung der Ausgangsdatei, die ausschließlich vollständige Datensätze (ohne Ausfallwerte) besitzt, erfolgt in drei Teildateien. In einer Teildatei werden die binären, in einer zweiten die stetigen Analysevariablen untergebracht. Die Identifikationsvariable wird gesondert abgelegt.
Das Makro arbeitet wie folgt:

1. Für die binären Merkmale werden bezüglich des Jaccard-Abstandes die Abstandsmatrix DISTBIN als Dreiecksmatrix erzeugt und die ersten acht Datensätze der Abstandsmatrix ausgegeben. Die ausgegebenen Abstände müssen bei richtiger Arbeitsweise im Intervall [0, 1] liegen.

2. Für die stetigen Merkmale werden die auf dem Mahalanobisabstand beruhende Abstandsmatrix erzeugt und der Normierungsschritt durchgeführt, sodass auch hier bei richtiger Arbeitsweise die ausgegebenen Abstände zwischen 0 und 1 liegen müssen. Dazu erfolgt zunächst der Aufruf der Prozedur CORR, von der die Kovarianzmatrix in einer Output-Datei COVMAT ausgegeben wird. In der anschließenden Prozedur IML werden die Matrix invertiert, die Mahalanobisabstände bestimmt und in der Datei DISTMAHAL abgespeichert.
3. Die Abstandsmatrizen DISTMAHAL und DISTBIN werden zur SUMABST Gesamtabstandsmatrix addiert. Die ausgegebenen Abstände müssen im Intervall [0, 2] liegen.

4. Die Prozedur CLUSTER wird aufgerufen. Die Abstandsmatrix SUMABST ist Eingabedatei. Im Output ist die Historie der Clusterung nachvollziebar. Durch die Option

METHOD=centroid pseudo ist im Programm die Clustermethode ausgewählt. Bei der Wahl einer anderen Methode sollte der Anwender darauf achten, daß auch alle weiteren notwendigen Parameter für die Analyse gesetzt sind.

5. Die Prozedur TREE illustriert die Clusterhistorie.

6. In einem letzten Output sind die zu einem Cluster gehörenden Individuen aufgelistet.

Das vollständige Programm ist im Anhang aufgeführt.

5. Ergebnisse

Ziel des Projektes ist die Klassifikation der Alkoholkonsumenten hinsichtlich ihres Gefährdungsgrades durch eine Kombination von Laborparametern und zusätzlicher Ja-Nein-Informationen. Wenn die sichere Einteilung in die Gruppen “trinkende Autofahrer” und “autofahrende Trinker” möglich ist, können die Indikatoren im Rahmen von Führerschein- oder Alkoholberatung und zur Anordnung einer weiterführenden Alkoholismusdiagnostik eingesetzt werden.
Von nur 968 Blutproben konnte der vollständige Datensatz erhoben und in die Clusteranalyse einbezogen werden, weil fehlende Daten eine Abstandsbestimmung nicht erlauben.
Bei der Betrachtung der Baumstruktur fällt auf, dass bei einer Distanz über 0.6824 ein sehr großes Cluster entsteht, das als einziges in den weiteren Prozedurschritten wächst. Deshalb wurde bei diesem Abstand, der einer Clusteranzahl von 15 entspricht, abgebrochen. Es entstehen vier größere und elf kleinere Cluster, von denen 7 weniger als 10 Personen enthalten. Von den größeren Clustern lassen CL1 (n=455) und CL2 (n=177) eine Deutung im obigen Sinne zu, weil sich ihre Merkmalsausprägungen gegensätzlich verhalten.
In Tabelle 3 ist zusammengestellt, wie viele der drei untersuchten Laborparameter bei den zu den Clustern CL1 und CL2 gehörenden Personen erhöht ist. Dieses deutet bekanntlich auf Alkoholismus hin. Deutlich ist zu erkennen, daß von 455 Personen 347 (76.3%) höchstens einen Parameter erhöht haben. Bei CL2 sind das dagegen nur 55 von 177, also 31.1%. Ebenso deutlich fallen die Unterschiede beider Cluster beim Abnahmezeitpunkt Tag/Nacht (Tabelle 4), Wochenende/Woche (Tabelle 5) und bei der Altersverteilung (Tabelle 6) auf.

	kein Parameter erhöht	1 Parameter erhöht	2 Parameter erhöht	3 Parameter erhöht	Summe
Cluster 1	195	152	65	43	455
Cluster 2	17	38	61	61	177
Rest	127	122	60	27	336
Summe	339	312	186	131	968

Tab.3: Anzahl von Personen mit erhöhten, Alkoholmissbrauch anzeigenden Laborparametern

	Nacht	Tag	Summe
Cluster 1	359	96	455
Cluster 2	0	177	177
Rest	319	17	336
Summe	678	290	968

Tab.4: Anzahl von Personen, die am Tag bzw. in der Nacht auffällig waren

	Woche	Wochenende	Summe
Cluster 1	0	455	455
Cluster 2	172	5	177
Rest	279	57	336
Summe	451	517	968

Tab. 5: Anzahl von Personen, die am Wochenende bzw. in der Woche auffällig waren

	Alter < 30	Alter ≥ 30	Summe
Cluster 1	233	222	455
Cluster 2	24	153	177
Rest	137	199	336
Summe	394	574	968

Tab. 6: Altersverteilung in den Clustern

Zusammenfassend kann der trinkende Autofahrer (autofahrende Trinker) archetypisch charakterisiert werden:

1. Er ist unter (mindestens) 30 Jahre alt.

2. Er wird am Wochenende nachts (in der Woche tagsüber) im Straßenverkehr auffällig.

3. Bei ihm ist (sind) höchstens einer (mindestens zwei) der Alkoholismus anzeigenden Laborparameter GGT, CDT und Methanol erhöht.

6. Literatur

Aderjan, R.: Marker missbräuchlichen Alkoholkonsums, Wissenschaftliche Verlagsgesellschaft Stuttgart (2000)

Falk, M., Becker, R., Marohn, F.: Angewandte Statistik mit SAS, Springer Verlag Berlin, Heidelberg New York (1995)

Hartung, J., Epelt, B.: Multivariate Statistik, R. Oldenbourg Verlag München Wien (1995)

SAS Institute Inc., SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 1, Cary, NC: SAS Institute Inc., 1989

Schuemer, R., Ströhlein, G., Gogolok, J.: Datenauswertung und statistische Analyseverfahren, G. Fischer Verlag Stuttgart New York (1995)

Späth, H.: Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion, Oldenbourg Verlag –2.Aufl.- München (1977)

7. Anhang

```
*************************************************;
*************************************************;
**                                             **;
**  Programm zur Durchfuehrung einer  Cluster-  **;
**  Analyse  mit binaeren und stetigen Merkma-  **;
**  len. An das Macro uebergeben  werden drei   **;
**  Dateien:                                    **;
**  a)Datei mit genau den alternativen Merkma-  **;
**    len als NUMERISCHE Felder mit 0 oder 1    **;
**  b)Datei mit genau den stetigen Merkmalen.   **;
**  c)Eine Datei, die NUR die ID-Variable ent-  **;
**    haelt                                     **;
**  Anz ist die Anzahl der Cluster,bei der das  **;
**  Verfahren stoppen soll. IDName ist der      **;
**  SAS-Name der Identifikationsvariablen       **;
**                                             **;
*************************************************;
*************************************************;

%MACRO CLUSTERMIX(Binaer,Stetig,Ident,IDName,Anz);

        ******************************;
        *       Abstandsmatrix       *;
        *    fuer binaere Daten      *;
        * Vergleiche Programm 7_2_4   *;
        * Falk, Becker, Marohn (1995) *;
        ******************************;
 PROC IML;
   USE &Binaer ;
   READ ALL VAR _NUM_ INTO x;
   p=NCOL(x);    * p = Anzahl der Merkmale;
   n=NROW(x);    * n = Anzahl der Beobachtungen;
   ** Berechnung der Distanz-Matrizen ;
   ** als untere Dreiecksmatrizen ;
   ** Koeffizienten durch Entfernen des Kommentars auswaehlen;
   dist=J(n,n,.);
   DO k=1 TO n;
      DO l=1 TO k;
         a=x[k,]*x[l,]`;
         b=x[k,]*(1-x[l,]`);
         c=x[l,]*(1-x[k,]`);
         d=(1-x[k,])*(1-x[l,]`);
         IF d=p THEN dist[k,l]=0; ELSE
         dist[k,l]=1 - a/(a+b+c);      * Jaccard; */
   *     dist[k,l]=1 - 2*a/(2*a+b+c);  * Czekanowski;
   *     dist[k,l]=1 - (a+d)/p;        * Matching;
      END;
   END;
```

```
   CREATE distbin FROM dist;  * Matrix => SAS-Datei;
   APPEND FROM dist;
 QUIT; RUN;                   * IML wird verlassen;


**   Erzeugen der SAS-Datei zum Druck der binären Abstände;
 DATA distb;
   MERGE &IDENT distbin;      * Erste Datei ist ID;
 RUN;
 TITLE1 'Auszug aus der Distanzsmatrix zu binaeren Merkmalen';
 PROC PRINT DATA=distb(OBS=8) ROUND;
   VAR COL1-COL8; ID  &IDName;
 RUN;
         ********************************;
         *   Distance-Matrix mit dem    *;
         *     MAHALANOBIS-Abstand      *;
         * Vergleiche Programm 7_2_2    *;
         * Falk, Becker, Marohn (1995)  *;
         ********************************;
 PROC CORR DATA=&STETIG OUTP=covmat COV NOPRINT;
 RUN;  * Mit allen stetigen Variablen;
 PROC IML;
   ** Einlesen der Dateien &STETIG und covmat;
   USE &STETIG;
   READ ALL VAR _NUM_ INTO x;
   n=NROW(x);     * n = Anzahl der Beobachtungen;
   USE covmat WHERE(_TYPE_='COV');
   READ ALL VAR _NUM_ INTO s;
   **  Berechnung der Distanz-Matrix dist als untere Dreiecksmatrix;
   dist=J(n,n,.);
   s_inv=INV(s);  * s_inv=Inverse von s;
   DO k=1 TO n;
      DO l=1 TO k;
      d1=(x[k,]-x[l,]);
      dist[k,l]=SQRT(d1*s_inv*d1`); * Mahalanobis-Abst.;
      dist[k,l]=dist[k,l]/(1+dist[k,l]); * Normierung auf [0,1);
      END;
   END;
   CREATE distmahal FROM dist;  * Matrix => SAS-Datei;
   APPEND FROM dist;
 QUIT; RUN;  * IML verlassen;

 **   Erzeugen der SAS-Datei zum Druck der normierten Mahalanobis-Abstaende;
 DATA mahal;
   MERGE &IDENT  distmahal; * Erste Datei ist ID;
 RUN;
 TITLE1 'Auszug aus der Mahalanobis-Distanzmatrix';
 PROC PRINT DATA=mahal(OBS=8) ROUND;
   VAR COL1-COL8; ID  &IDName;
 RUN;
         ********************************;
         * Distance-Matrix mit der Summe *;
         * aus  binaerem und MAHALANOBIS *;
         ********************************;
 PROC IML;
   USE distmahal;
   READ ALL VAR _NUM_ into x;
   n=NROW(x);
   USE distbin;
   READ ALL VAR _NUM_ into y;
   gem=J(n,n,.);
   DO k=1 to n; gem[k,]=x[k,]+y[k,];
   END;
   CREATE SUMABST FROM gem;
   APPEND FROM gem;
 QUIT; RUN;
**   Erzeugen der SAS-Datei zum Druck des Summen-Abstandes;
 DATA sumabst;
   MERGE sumabst &IDENT; * zweite Datei ist ID;
 RUN;
 TITLE1 'Auszug aus der Summen-Distanzmatrix';
 PROC PRINT data=sumabst(OBS=8)ROUND;
   VAR COL1-COL8; ID  &IDName;
 RUN;
 TITLE1 'Clusteranalyse mit gemixten Daten';
 PROC CLUSTER data=sumabst(type=distance)
   METHOD =centroid pseudo OUTTREE=tree;
```

```
   ID &IDName;
 RUN;
 PROC TREE data=tree N=&Anz OUT=out;
   ID &IDName;
 RUN;
** Ausgabe der ID bezueglich der gefundenen Cluster;
 PROC SORT DATA=out;
   BY CLUSTER;
 RUN;
 PROC PRINT NOOBS;
   VAR &IDName; BY cluster;
 RUN;
%MEND CLUSTERMIX;

%CLUSTERMIX(sasuser.alkbin,sasuser.alksteti,sasuser.alkid,lfdnr,15);

RUN;
```