Über die multivariate Aufdeckung von Ausreißern in Kohortenstudien

Hans-Peter Altenburg
Deutsches Krebsforschungszentrum
Klinische Epidemiologie
69120 Heidelberg
hp.altenburg@dkfz.de

Zusammenfassung

Es wird der so genannte Epidemie-Algorithmus und seine Realisierung mit Hilfe des SAS-Systems beschrieben. Die dem Epidemie-Algorithmus zu Grunde liegende Idee (Hulliger und Béguin (2001)) ist, einen (einfachen) Epidemie-Prozess über die Hauptmasse der Daten zu legen. Ausgangsbasis des Prozesses ist ein Punkt (ein „Infektiöser“) in oder zumindest sehr nahe der Mitte dieser Masse. Alle anderen Punkte sind zu Beginn „Suszeptible“, d.h. nicht infiziert, die aber angesteckt werden können. Die Epidemie breitet sich über die gesamte Datenmasse aus, und alle Datenpunkte werden nach und nach bzw. innerhalb eines bestimmten vorgegebenen Zeitintervalls infiziert. Der Prozess der Identifizierung eines Ausreißers ist dann stark mit der Zeitdauer bis zur Infektion eines Datenpunktes assoziiert. Mit anderen Worten: Die Infektionszeit, eine zufällige Abbildung der Population (d.h. des $\Re^n$) auf die Zeitachse (=$\Re^1$), wird als Kriterium zur Ausreißeridentifikation benutzt: „Ausreißer“ haben eine besonders lange Infektionszeit. Der Prozess besteht im Wesentlichen aus zwei Komponenten: ein grundlegendes Abstandsmaß, welches in die Infektionswahrscheinlichkeit eingeht, und ein geeignetes Modell für den Epidemieprozess. Beide Komponenten müssen dem Datentyp entsprechend angepasst werden.

Keywords: Ausreißererkennung, Epidemiealgorithmus, Ähnlichkeitsmaße, Distanzmaße.

1 Einleitung

Große epidemiologische Kohortenstudien mit einer Vielzahl von Probanden und auch Variablen haben Ausreißer und ihre Behandlung als beständiges

Problem. In der Regel ist die Erkennung von Ausreißern wichtig, um die Datenqualität zu verbessern. Die Studien-Evaluation, d.h. die Kontrolle und Korrektur der Daten, ist ein kostenintensiver Prozeß, wobei auch die Imputation schlechter Daten oder von Ausreissern sorgfältig geplant werden muß. Insbesondere erscheint es wünschenswert, Fälle nicht einfach deshalb ausschließen zu müssen, weil einige der Kovariablen besonders extreme Werte aufweisen. Ausreißererkennung wird schon einfach deshalb benötigt, um die Datenqualität zu verbessern. Das Problem stellt sich besonders auch als multivariate Fragestellung. Die Motivation für die Beschäftigung mit der Problemstellung kommt von der Analyse epidemiologischer Daten der EPIC-Studie (**E**uropean **P**rospective **I**nvestigation into **C**ancer und Nutrition), einer Kohortenstudie mit neun beteiligten Ländern und 29 Zentren, wobei in die Auswertung ca. 500000 Probanden und je nach Fragestellung 300 bis 900 Variablen eingehen.

Ein zentrales Problem stellt auch die variierende Datenqualität dar, einmal bei den Daten der Ernährungsaufnahme (Variablen mit Messfehlern) sowie unzureichend gemessene Daten, wie z.B. beim Rauchen (Intensität, Dauer, Zeit seit Ende Rauchen, ...), richtig gemessene Daten, wie etwa Körpergewicht und Größe, sowie auch nicht normal-verteilte Daten. Eine heterogene Daten-Qualität existiert sowohl zwischen den Ländern als auch zwischen den Zentren.

Im Folgenden soll der so genannte Epidemie-Algorithmus und seine Realisierung beschrieben werden. Die dem Epidemie-Algorithmus zu Grunde liegende Idee (Hulliger und Béguin (2001)) ist, einen (einfachen) Epidemie-Prozess über die Hauptmasse der Daten zu legen und sich dort ausbreiten zu lassen:

- Starte eine einfache Epidemie (engl.: simple epidemic process) von einem geeignet gewählten zentralen Punkt aus (z.B. der räumliche Median mit Hilfe der SAS-Prozedur `MEANS` und der Option `Median`).
- Am Anfang gibt es ein infiziertes Objekt, d.h. einen „Infizierten“ (I) und *n-1* mögliche Empfänger, sog. Suszeptible (S) Datenobjekte.
- Die Epidemie breitet sich dann über die gesamte „Population“ (Datenpunkte, Stichprobe) aus, wobei die meisten Datenpunkte infiziert werden (zu „Infizierten“ werden).
- Der Prozeß der Ausreißer-Identifikation ist dann stark verwandt mit der Zeitdauer, wann ein Datenpunkt (Suszeptibler) infiziert wird (d.h. eine zufällige Abbildung $\Re^n \rightarrow \Re^1$).

SAS-Programm-Statements für die Bestimmung des Startpunktes:

```
/* Bestimmung der Koordinaten des Startpunktes */
PROC MEANS DATA=&dset MEAN MEDIAN ;
VAR &varlist ;
RUN ;

/* Variante mit Ausgabe in eine Datei "dsum"
   mit Hilfe von ODS */
ODS OUTPUT SUMMARY=dsum ;
PROC MEANS DATA=&dset MEAN MEDIAN ;
VAR &varlist ;
RUN ;
ODS OUTPUT CLOSE ;
```

Für die Realisierung des Algorithmus werden im wesentlichen zwei Komponenten benötigt, die genauer spezifiziert werden müssen:

- Ein grundlegendes Abstandsmaß, welches in die Infektionswahrscheinlichkeit eingeht bzw. die Zeit bis zu einer Infizierung beeinflusst, und
- ein geeignetes Modell für den Epidemieprozess, wie z.B. stetiges oder diskretes Zeitmodell.

Beide Komponenten müssen dem Typ der Daten entsprechend angepasst werden, was u.U. eine nichttriviale Aufgabe sein kann.

2 Distanz- oder Ähnlichkeitsmaße

In einem ersten Schritte muß für alle Datenpunkte die Distanz zu allen anderen Datenpunkten ermittelt werden, damit mit deren Hilfe die Infektionszeiten bestimmt werden können. Hierbei wird die Beobachtungsmatrix X in eine Distanz- (Dreiecks-) Matrix D transformiert, z.B.

- Euklidischer Abstand (nur für stetige Variablen)
$$d_{ij} = [\Sigma_k (x_{ik} - x_{jk})]^{1/2}$$
- Jaccard-Koeffizienten
- „Gower‘s general coefficient“ (diskrete oder dichotome Variablen, siehe Gower (1971), Gower und Legendre (1986) oder Hubálek (1982)), ...

Es existieren für Distanzmaße sehr viele mögliche Varianten, auf die hier nicht weiter eingegangen werden kann (siehe hierzu Hubálek (1982) oder Jäger et al. (2001)). Ein vorsichtiges Vorgehen für nicht-metrische Skalen oder Beobachtungsgrößen mit eingeschränkter Variabilität (z.B. Ernährung-Variablen) ist angebracht. Zur Vermeidung unbalanzierter Effekte bei den verschiedenen Variablen ist es u.U. besser, standardisierte Varianten der Variablen zu verwenden, wobei z.B. Streuungskennzahlen mit Hilfe robuster Varianz-Maße (z.B. MAD, vgl. Rousseeuw (1987)) bestimmt werden sollten.
Die Realisierung der Distanzbestimmung kann im SAS-System mit Hilfe des SAS-Macros `%DISTANCE` erfolgen. In diesem Macro sind bereits eine große Zahl von Abstandsmaßen verwirklicht, so dass nur ein geringer Programmieraufwand erforderlich ist.

Aufruf des SAS-Makros DISTANCE:

```
/* Bestimmung der Distanz zwischen den
   Punkten mit Hilfe des Macros %DISTANCE */
%DISTANCE(DATA=&dset,
          VAR=&varlist,
          ID=&idvar,
          OUT=&outset,
          METHOD=EUCLID,  ... )
```

Zulässige Variablentypen sind dabei**:**

R	Ratio — numerisch
I	Interval — numerisch
O	Ordinal — numerisch
N	Nominal — numerisch oder Character

Je nach Variablentyp müssen unterschiedliche Distanzkennzahlen verwendet werden, die insbesonders bei diskreten Messgrößen nur eine eingeschränkte Variabilität aufweisen können. Das Distanzmakro erlaubt zahlreiche Methoden der Distanzbestimmung. Die nachfolgende Liste enthält zulässige Schlüsselworte für die Bestimmung einer Distanz, z.B.

GOWER	Gower's Ähnlichkeit
DGOWER	1-Gower
EUCLID	Euklid-Abstand
SQEUCLID	quadrierter Euklid-Abstand
SIZE	Size Distanz
SHAPE	Shape Distanz

COV	Kovarianz
CORR	Korrelation
DCORR	Korrelation transformiert in Euklid-Abstand als $\sqrt{}$(1-CORR)
SQCORR	quadratische Korrelation
DSQCORR	1 – (quadratische Korrelation)
L(p)	Minkowski L_p-Distanz, $p>0$
CITY	L_1-Norm
CHEBYCHE	L_∞-Norm
POWER(p,r)	Verallgem. Euklid-Abstand, $p>0$, $r\geq 0$

3 Epidemie-Prozess („Simple Epidemic“)

Erste einfache Modelle zu Epidemien wurden bereits zu Beginn des letzten Jahrhunderts entwickelt, z.B. von Hamer im Jahr 1906. Im Prinzip hängt der Verlauf einer „einfachen“ Epidemie dabei ab von der Anzahl der Suszeptiblen, der Anzahl der Infizierten und der Kontakt-Rate. Die Modelle von Hamer waren dabei noch deterministische Modelle. 1926 wurden grundlegende diskrete probabilistische Modelle von McKendrick entwickelt. Im Jahr 1949 nach dem zweiten Weltkrieg modellierte Bartlett einfache, stetige Epidemien, die sehr ähnlich zu Geburts- und Todesprozessen waren. Für unseren Algorithmus wird nur ein sehr einfaches Epidemie-Modell benötigt:

$$S \rightarrow I,$$

d.h. es gibt nur zwei Gruppen von Objekten, Suszeptible und Infizierte. Suszeptible können in den Zustand „infiziert“ übergehen, Infizierte können sich nicht nicht mehr erholen, sondern bleiben während des gesamten Prozesses infiziert. Ein Infizierter Datenpunkt kann die Erkrankung beliebig lang übertragen. Ziel eines Modells ist es, die Verteilung der Anzahl der Infizierten (# I) in einer Population zu bestimmen. Als Modellparameter wird nur die Infektions-Rate benötigt.

Start-Punkt ist der „sample spatial“ Median (Stichprobenpunkt in der “Mitte“ der Datenmasse, eine Minimaleigenschaft charakterisierend). Als Zeitskala des Epidemie-Prozesses können sowohl diskrete Zeit als auch stetige Zeit gewählt werden. Die Infektionsrate geht ein in die Übertragungswahrscheinlichkeit für die Infektion. Der Infektionsprozess verläuft also etwa in der Form: Gegeben ein Infizierter ist im Punkt *i*, so ist die Wahrscheinlichkeit, daß ein Nicht-Infizierter in Punkt *j* infiziert wird, eine Funktion *h* des Abstands zwischen *i* und *j*. Sie hängt also ab von der Distanz

eines Infizierten von einem Suszeptiblen, und sollte mit der Entfernung zwischen den beiden Punkten fallen. Beispiele für diese Transmissions-Funktionen sind:

- Lineare Funktion
- Logistische Verteilung (Logit),
- Normal-Verteilung (Probit),
- Weibull-Verteilung,
- Pearson-Verteilung
- Kontakt-Verteilungen aus „Epidemics“

Im Prinzip kann jede Wahrscheinlichkeitsverteilung verwendet werden.
Je nach Situation der Daten lassen sich auch alternative Transmissionsprozesse formulieren, die dann entsprechend andere Verteilungen implizieren: Folgt beispielsweise die Infektion in jedem Zeitintervall einem zufälligen Poisson-Prozess, so ist die Verteilung des Zeitintervalles bis zum nächsten Ereignis negativ exponential verteilt.
Programmschritte für ein stetiges Zeitmodell für die Erzeugung einer exponential erzeugten Zufallszahl im eindimensionalen Fall:

```
d2=ABS(x-LAG(x)) ;
time=RANEXP(seed)*(d1) ;
```

Dabei ist `x` die (eindimensionale) Messgröße und `d2` der Euklidische Abstand zum letzten infizierten Punkt. Die zweite Zeile bestimmt eine exponentiell verteilte Zeitdauer, welche die Zeit darstellt, die vergeht, bis der nächste Punkt infiziert wird.

4 Algorithmus

Im Modell mit diskreter Zeit läßt sich der Algorithmus wie folgt beschreiben (vgl. Hulliger and Béguin (2001)):

1. Setze die Infektions-Zeit aller Punkte auf null.
2. Wähle den Stichprobenmedian als Start-Punkt und setze seine Zeit $t_{ssm}=1$.
3. Erhöhe die Infektionszeit um *1*.
4. Berechne die totale Infektions-Wahrscheinlichkeit p für alle nichtinfizierten Datenpunkte ("suszeptible").

5. Wähle den nächsten infizierten Punkt aus mit Hilfe eines Bernoulli-Prozesses mit Erfolgs-Wahrscheinlichkeit p. Setze seine Infektions-Zeit auf t.
6. Update der Menge der infizierten Datenpunkte.
7. Stopp, wenn alle Punkte infiziert sind, ansonsten fahre mit Punkt 3 fort.

Der Algorithmus stoppt, wenn alle Datenpunkte infiziert sind oder keine Infektion in einem vorgegebenen Zeitintervall der Länge l geschieht, was bei einem diskreten Zeitmodell passieren kann (deshalb werden stetige Zeitmodelle empfohlen!). Nichtinfizierte Datenpunkte behalten dann die Infektions-Zeit *0*. Ein Problem dabei ist: Die Zahl l muss im Voraus bestimmt werden.
Alle Schritte des Algorithmus können im Prinzip über ein Makro, was DATA-Step sowie Sortiervorgänge ausführt, abgewickelt werden. Insbesondere kann dabei die Auswahl des nächsten infizierten Punktes in einem Bernoulli-Prozess in einem DATA-Step mit Hilfe der Funktion RANUNI bzw. UNIFORM erfolgen.

5 „Computational“ Aspekte

Im Fall vieler Datenpunkte kann die Berechnung der Distanzmaße zeitraubend sein (Modell-abhängig). Für die Realisierung des Algorithmus wird eine detaillierte Diskussion der Infektions-Wahrscheinlichkeit nach der Daten-Situation empfohlen, was natürlich die Praktikabilität des Algorithmus stark einschränkt. Problematisch kann auch die optimale Wahl der Infektions-Wahrscheinlichkeits-Funktion werden. Der Algorithmus ist verwandt zu den „data depth“-Methoden, verlangt aber nicht elliptische Verteilungen. Das Stopp-Kriteriun (Zahl l) im diskreten Zeitmodell ist abhängig von der Wahrscheinlichkeit für keine Infektion in l Versuchen. Weitere Erfahrungen mit dem Algorithmus müssen noch gesammelt werden, um Eigenschaften der Ausreißererkennung, wie z.B. Breakdown Point, herauszufinden.
Der Algorithmus konvergiert langsamer als andere effektivere „deterministische“ Verfahren der Ausreißererkennung. Vorteilteilhaft ist die geringe Berechnungszeit. Die Anzahl der Versuche bis zum Abbruch ist polynomial verteilt. Sie hängt nicht ab von der Dimension p der Variablen, die betrachtet werden sollen. p beeinflusst lediglich die Berechnung der Distanz. Ein Vorteil des Verfahrens ist auch, dass keine Transformationen und keine Vor-

auswahl (gute / schlechte Daten) erforderlich sind. Nachteilig wirken sich Verteilungen mit geringer Variabilität z.B. Ernährungs-Variablen aus Fragebogenerhebungen aus.

6 Schlussfolgerungen

Multivariate Ausreisser sind meist heterogen. Eine allgemeine Erkennungsmethode muss deshalb nicht unbedingt gut funktionieren, um alle Ausreisser zu finden. Ausreisser hinsichtlich der Lage müssen nicht notwendig Ausreisser bzgl. einer Korrelations- Struktur sein! Speziell für Ernährungsdaten, die oft nur eine eingeschränkte Variabilität aufweisen, kann dieser Ausreisser-Algorithmus schief gehen. Für Alternativen oder Verbesserungen des Algorithmus sei auf das Buch *Theory of Epidemics* von Bailey (Bailey 1975) und die erwähnte umfangreiche Literatur verwiesen. Es kann insbesondere dann hilfreich sein, wenn etwa die diskrete Zeitskala und der zugeordnete Bernoulli-Prozess in einen stetigen Zeit- bzw. Kontakt-Prozess übergeführt werden sollen.

Literatur

[1] Bailey, N.T.J. (1975): The Mathematical Theory of Infectious Diseases. Griffin, London

[2] Hulliger, B. and Béguin, C. (2001): Detection of multivariate outliers by a simulated epidemic. Eurstat, 667-676

[3] Gower, J.C. (1971): A general coefficient of similarity and some of its properties. Biometrics 27, 857-871

[4] Gower, J.C., Legendre, P. (1986): Metric and Euclidian properties of dissimilarity coefficients. Journal of Classification 3, 5-48

[5] Hubálek, Z. (1982): Coefficients of association and similarity based on binary (presence-absence data): an evaluation. Biological Reviews, 57, 669-689

[6] Jäger B., Wodny M., Rudolph P.E., Patschinsky, D. (2001): Clusteranalyse mit Binärdaten. In: M. Schuhmacher et al.: Proceedings 5. KSFE, Universität Hohenheim, 8.-9. März 2001

[7] Rousseeuw, P.J. and Leroy, A.M. (1987): Robust Regression and Outlier Detection. New York, Wiley