Diskriminanzanalyse mit binären Daten

Bernd Jäger, Michael Wodny, Karl-Ernst Biebler
Institut für Biometrie und Medizinische Informatik
E.-M.-Arndt-Universität
W.-Rathenau-Str. 48
17487 Greifswald
bjaeger@biometrie.uni-greifswald.de
wodny@biometrie.uni-greifswald.de
biebler@biometrie.uni-greifswald.de

Paul Eberhard Rudolph
Forschungsinstitut für die Biologie
landwirtschaftlicher Nutztiere
W.-Stahl-Allee 2
18196 Dummerstorf
Rudolph@fbn-dummerstorf.de

Karen Matthies
Klinik und Poliklinik für Innere Medizin A
Friedrich-Löffler-Str. 23 a
17487 Greifswald
kraatz@uni-greifswald.de

Zusammenfassung

Diskriminanzanalysen liegen in Statistikprogrammen für stetige Merkmale vor. Sind kategoriale Daten auszuwerten, behilft man sich oft mit Stetigkeitstransformationen oder durch die Annahme, die vorliegenden Daten sind eingeschränkte Beobachtungen stetiger Merkmale.
In der vorliegenden Arbeit wird ein Diskriminationsproblem für zwei Gruppen und Binärdaten mit 0-1-Codierung behandelt. Diese Diskriminanzanalyse beruht auf einer variierten n-nächste-Nachbarn-Methode und nimmt Bezug auf die simple matching-Metrik bzw. die Tanimoto-

Metrik. Neben der Lösung des Diskriminationsproblems wird ein Verfahren zur Merkmalsreduktion vorgestellt. Es handelt sich dabei um ein Ein-Schritt-Verfahren unter Bezug auf die Lachenbruch-Reklassifikation. Die rechentechnische Realisierung erfolgte in einer SAS®-Umgebung als IML-Programm. Eine Anwendung auf Daten aus der medizinischen Forschung demonstriert die Verfahren.
Die Vorgehensweise kann prinzipiell auf beliebige kategoriale Daten mit einer vorgegebenen Codierung sowie für den Fall mehrerer Gruppen verallgemeinert werden.

Keywords: nonparametric classification, binary variables, discriminant analysis, n-nearest-neighborhood method.

1 Einleitung

Wird ein Individuum durch einen r-dimensionalen reellen Vektor x charakterisiert, $x = (x_1, x_2, \ldots, x_r) \in \mathbb{R}^r$, so finden üblicherweise als Abstand zweier solcher Individuen bzw. Vektoren x und y die Euklidische Metrik

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

oder die Mahalanobis-Metrik

$$d_M(x, y) = (x - y) COV^{-1} (x - y)^T$$

Verwendung in Klassifikationsverfahren. Hier bezeichnet COV^{-1} die Inverse der empirischen Kovarianzmatrix der Daten.

Sind zwei Klassen (Gruppen) von Individuen gegeben, so lässt sich das Zuordnungsproblem (Diskriminanzproblem) für ein neu zu klassifizierendes Individuum z beispielsweise durch die n-nächste-Nachbarn-Regel lösen. Dazu werden die n bzgl. einer gewählten Metrik nächsten Nachbarn von z aus der Datenmenge herausgesucht. Gehören davon mehr als $n/2$ zur Klasse 1, wird z dieser zugeordnet, ansonsten gruppiert man z zur Klasse 2. Modifizieren lässt sich diese Zuordnungsregel durch Wichtung der Abstände, etwa mittels des relativen Anteils der jeweiligen Gruppe an der gesamten Datenmenge.

2 Abstände für Binärdaten

In der medizinischen Forschung besonders häufig anzutreffen ist die Kennzeichnung des Gesundheitszustandes eines Individuums durch das Vorhandensein

bzw. Nichtvorhandensein von r Symptomen. Einem Individuum entspricht damit ein Wort der Länge r, bestehend aus den Ziffern 1 (Vorhandensein des Symptoms) und 0 (Nichtvorhandensein des Symptoms).

Für solche binären Vektoren ist die sogenannte „simple matching distance“ d_{SM} ein möglicher Abstand. Die Definition dieser Metrik soll an einem Beispiel demonstriert werden:

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

gibt die Tabelle 1 an, dass in $\alpha = 2$ Koordinaten sowohl x als auch y den Wert 1 haben. Für $\beta = 1$ Koordinaten hat x den Wert 1 und y den Wert 0, für $\gamma = 1$ Koordinaten hat x den Wert 0 und y den Wert 1, für $\delta = 4$ Koordinaten hat sowohl x als auch y den Wert 0. Damit zählen $\alpha + \delta$ die übereinstimmenden und $\beta + \gamma$ die nicht übereinstimmenden Koordinaten.

Es ist

$$d_{SM}(x,y) = 1 - \frac{\alpha + \delta}{n}$$

eine Metrik auf $\{0,1\}^8$. Stimmen x und y überein, ergeben sich $\alpha + \beta = n$ und somit $d_{SM}(x,x) = 0$. Für obige x und y gilt

$$d_{SM}(x,y) = 1 - \frac{2+4}{8} = \frac{1}{4} \quad .$$

Tabelle 1: Anzahlen $\alpha, \beta, \gamma, \delta$ der in den angegebenen Vektoren x und y beobachteten 4 verschiedenen Paare (x_i, y_i).

		y_i		
		1	0	$\sum$
x_i	1	$\alpha = 2$	$\beta = 1$	3
	0	$\gamma = 1$	$\delta = 4$	5
	$\sum$	3	5	$r = 8$

Als weiterer Abstandsbegriff, der im vorgestellten SAS-Programm wahlweise verwendet werden kann, wurde die Jaccard-Metrik (auch Tanimoto-Metrik) gewählt,

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

Sie ermöglicht differenziertere Abstandsbewertungen als d_{SM}, denn beispielsweise für $z = (0,0,1,0,0,1,1,1) \in \{0,1\}^8$ gelten $d_{SM}(x,y) = d_{SM}(x,z) = 3/8$, aber $d_J(x,y) = 3/5$ und $d_J(x,z) = 1/2$.

3 Variierte n-nächste-Nachbarn-Methode

Für Daten beobachteter stetiger Merkmale sind die Wertemengen der Euklidischen Metrik bzw. der Mahalanobis-Metrik überabzählbar. Der nächste Nachbar eines zu klassifizierenden Individuums ist eindeutig bestimmt.
Die Wertemengen der Metriken d_{SM} und d_J für Binärdaten sind nicht sehr umfangreich. Bezüglich d_{SM} sind dies für Vektoren der Länge r die $r+1$ verschiedenen Werte $0, 1/r, 2/r, \ldots, (r-1)/r, 1$. Wird die Euklidische Abstandsmessung anwendet, ergeben sich genauso viele, jedoch andere Abstandswerte. Im $\mathbb{R}^3$ können die acht möglichen 0-1-Vektoren als Ecken eines Würfels in Ursprungslage aufgefasst werden. Die möglichen von 0 verschiedenen Abstände in der Euklidischen Metrik (entsprechend bzgl. d_{SM}) sind der Abstand 1 (entsprechend 1/3) zwischen benachbarten Eckpunkten, $\sqrt{2}$ (entsprechend 2/3) zwischen Eckpunkten auf einer Seitendiagonale und $\sqrt{3}$ (entsprechend 1) zwischen Eckpunkten auf der Raumdiagonale.

Bei der Klassifizierung bzgl. Binärdaten ist zu berücksichtigen, dass das zu klassifizierende Individuum zu mehreren Nachbarn den gleichen Abstand aufweisen könnte. Das veranlasst zu folgender Variation der n-nächste-Nachbarn-Methode:

A Ist die Zuordnung des Individuums z zu einer der beiden Klassen bereits bezüglich seiner $(n-1)$-ten Nachbarn möglich, wird die n-nächste-Nachbarn-Regel beibehalten.

B Trifft A nicht zu, so sind alle k Nachbarn gleichen Abstandes zu z, die als n-te Nachbarn gelten können, in die Entscheidung einzubeziehen. Die Zuordnung erfolgt in diesem Fall nach einer $(n+k)$-nächste-Nachbarn-Regel.

Während n für das Klassifizierungsverfahren vorab zu wählen ist, ergibt sich k aus den Daten. Diese Entscheidungsregel kann selbstverständlich durch Wichtungen modifiziert werden.

4 Merkmalsreduktion

Um den Einfluss der Anordnung der Merkmale auf das Ergebnis der Merkmalsreduktion zu minimieren, verfährt man vorbereitend wie folgt:
Jedes einzelne Merkmal wird anhand der assoziierten Vierfeldertafel dahingehend beurteilt, wie es allein die vorgegebene Klassifikation in der Lernstichprobe reproduziert. Die Spalten der Datei ordnet man entsprechend dieser Beurteilung.

Als Kriterium der Merkmalsselektion dient die Reklassifikation nach Lachenbruch. Nacheinander wird jedes Element der Lernstichprobe unter Bezug auf die verbleibenden Individuen nach der variierten n-nächste-Nachbarn-Methode klassifiziert. Von der so gewonnenen Reklassifikation der Lernstichprobe ist die Angabe des Anteiles der richtig klassifizierten Beobachtungen möglich. Er ist das Optimierungskriterium für die Merkmalsreduktion und soll maximal sein. Zunächst wird sukzessiv aus der Datei immer genau eine Variable entfernt und jeweils die oben beschriebene Reklassifikation durchgeführt. Der um die Variable reduzierte Variablensatz, bei dem die Reklassifikation am besten ausfällt, verbleibt im Verfahren. Sollten mehrere Variablensätze zu gleich guten Reklassifikationsresultaten führen, wird durch die anfangs durchgeführte Spaltenumordnung diejenige Variable gestrichen, die allein die vorgegebene Klassifikation am schlechtesten reproduziert.
Diese Merkmalsreduktion wird wiederholt.
Abbruchkriterien, etwa nach den richtig klassifizierten Datensätzen, sind nicht in das Verfahren eingebaut. Ohne Eingriff wird der Variablensatz so lange reduziert, bis er nur noch eine Variable enthält.
Die rechentechnische Realisierung erfolgte in einer SAS®-Umgebung als IML-Programm. Interessenten können sich an obige Anschrift wenden.

5 Anwendungsbeispiel

In einer Fall-Kontroll-Studie der Klinik und Poliklinik für Innere Medizin A (Direktor: Univ.-Prof. Dr. med. Günter Kraatz) des Universitätsklinikums Greifswald wurden an Morbus Wegner erkrankte Patienten untersucht. Dabei handelt es sich um besonders schwere Autoimmunerkrankungen, die zunächst die Atemwege und im fortgeschrittenen Stadium die Nieren so schwer schädigen, da die Patienten der Dialyse zugeführt werden müssen. Alle 30 Morbus Wegner Fälle wurden über Nierensprechstunden und Dialysezentren des Bundeslandes Mecklenburg-Vorpommern aufgefunden. Sie stellen wahrscheinlich die Krankheitspopulation des Bundeslandes dar.
Ausgewertet wurden Angaben zu 56 binären Variablen, 55 Datensätze waren vollständig. Die Variablen beziehen sich auf Erkrankungen in der Kindheit, Umgang mit Haustieren, Umweltgifte, aber auch auf Antikörperuntersuchungen bestimmter Krankheitserreger, auf Allergien und Allergieauslöser.
Es interessieren Wertigkeiten von Merkmalen und Merkmalskombinationen in Hinsicht auf die Erkrankung. Insbesondere sollte diejenige minimale Variablenmenge herausgefunden werden, die eine zufriedenstellende Klassifikation sowie eine medizinische Interpretation erlaubt. Das beschriebene Diskriminanzverfahren mit Merkmalsreduktion ergibt die in Tabelle 2 teilweise wiedergegebene Merkmalsreduktion.
Danach könnten die 56 Merkmale bis auf acht reduziert werden. Man erkennt

deutlich, da die Anzahl der Fehlklassifizierten mit derjenigen der verbleibenden Merkmale rasch absinkt, dann aber wieder ansteigt. Hier gilt es ein Optimum zu finden.
Ob die vorgeschlagene Auswahl von acht Variablen dem Anwender zur Interpretation sinnvoll erscheint oder ob aus medizinischer Sicht noch einige Merkmale hinzu zu nehmen sind, soll an dieser Stelle nicht erörtert werden.

Tabelle 2: Ausschnitt aus der Ergebnisliste des abbauenden Verfahrens der Diskriminanzanalyse mit Binärvariablen

Reduzierte Variable	Variablen-anzahl	Morbus Wegner		Kontrollgruppe	
		richtig	falsch	richtig	falsch
—	56	12	15	14	14
Neubau	55	14	13	19	9
Infanz	54	15	12	20	8
Myk_IgA	53	15	12	21	7
Blut_0	52	16	11	23	5
Land	51	18	9	24	4
CMV_IgM	50	19	8	24	4
Würmer	34	20	7	26	2
Org_Staub	33	20	7	26	2
Katze	32	20	7	26	2
Hund	31	21	6	26	2
Röteln	30	22	5	25	3
Blut_Rh	9	25	2	24	4
Blut_B	8	24	3	24	7
Tonsillekt	7	24	3	22	6
Säuger_sonst	6	24	3	21	7
Grippe	5	24	3	23	5
Metalle	4	22	5	25	3
Allergien	3	20	7	23	5
Myk_IgG	2	24	3	17	11

Literatur

[1] Asparoukhov, O.K. und Krzanowski, W.J. (2001). A comparison of discriminant procedures for binary variables. Comput. Stat. Data Anal. 38, 139-160.

[2] Asparoukhov, O.K. und Stam, A. (1997). Mathematical programming formulations for two-group classification with binary variables. Ann. Oper.Res. 74, 89-112.

[3] Hall, P. (1981). On nonparametric multivariate binary discrimination. Biometrika, 68, 287-294.

[4] Leung, C.-Y. (1994). Classification of dichotomous and continuous variables with incomplete samples. Commun. Stat., Theory Methods 23, No.6, 1581-1592.

[5] Vlachonikolis, I.G. (1986). On the estimation of the expected probability of misclassification in discriminant analysis with mixed binary and continuous variables. Comput. Math. Appl., Part A12, 187-195.

[6] Jäger, B., Wodny, M., Rudolph, P.E., Patschinsky, D. (2001). Clusteranalyse mit Binärdaten. Hohenheim, Stuttgart, 181-190.