Monatshefte für Chemie 107, 547—555 (1976)

Testverfahren für automatische Spektreninterpretationsmethoden

Von

K. Varmuza und **H. Rotter**

Institut für Allgemeine Chemie, Technische Universität Wien, Österreich

Mit 2 Abbildungen

(Eingegangen am 28. Oktober 1975)

Judgement of Automatic Spectra Interpretation Methods

Automatic spectra interpretation methods (e. g. pattern recognition methods for the interpretation of mass spectra) should be characterized by suitable criteria of quality. These criteria may be obtained by testing interpretation methods with a random sample of spectra, but should be independent from probabilities of classes in this sample, or should refer to a sample with equal probabilities of classes. In this paper, mathematical formulae for such objective criteria of quality are given. For example, individual predictive abilities for the classes and maximum information are appropriate to characterize and compare interpretation methods from different authors.

Einleitung

In den letzten Jahren wurden zahlreiche statistische und heuristische Methoden erprobt, um chemisch-physikalische Meßdaten mit Hilfe von Computerprogrammen automatisch interpretieren zu können. Besonders zahlreich waren die Anwendungsvorschläge für Methoden der automatischen Zeichenerkennung („pattern recognition") zur Ermittlung chemischer Partialstrukturen aus niedrig aufgelösten Massenspektren[1–3]. Derartige Spektreninterpretationsmethoden enthalten stets einen Klassifikator (Entscheidungsregel), der definiert, wie die Meßdaten (z. B. die Peakhöhen eines Massenspektrums) mathematisch zu verknüpfen sind; auf Grund des Ergebnisses dieser Rechnung wird dann das Spektrum einer bestimmten Klasse zugeordnet. Meist wird ein binärer Klassifikator verwendet, der das Spektrum einer von zwei einander ausschließenden Klassen zuordnet (z. B. Klasse 1: Partialstruktur x vorhanden, Klasse 2: Partialstruktur x nicht vorhanden). Die Klassifikatoren werden entweder mit statistischen Methoden aus dem

vorhandenen Datenmaterial (Spektrenbibliothek) abgeleitet oder nach bekannten chemisch-physikalischen Gesetzmäßigkeiten (z. B. Zerfallsregeln) entwickelt. Man erhält im allgemeinen Klassifikatoren, die richtige Antworten nicht zu 100% liefern.

Die Brauchbarkeit eines Klassifikators hängt vom Anwendungsfall ab und kann letzten Endes nur bei praktischen Einsätzen ermittelt werden. Dieser subjektiven Beurteilung sollte jedoch eine mathematisch-statistische Charakterisierung des Klassifikators vorausgehen. Zu diesem Zweck wird eine Stichprobe bekannter Spektren klassifiziert — meist sind es Spektren, die für die Entwicklung des Klassifikators nicht verwendet wurden, also dem Klassifikator „unbekannt" sind. Aus dem „Prozentsatz richtig klassifizierter Spektren" läßt sich die Güte des Klassifikators ableiten. Wie bereits früher[4] gezeigt wurde, sind die meisten der bisher in der Literatur verwendeten Güteangaben jedoch nicht geeignet, Klassifikatoren objektiv zu bewerten. Im folgenden werden die beim Test eines binären Klassifikators anfallenden Kenngrößen systematisch abgeleitet und auf ihre Brauchbarkeit zur objektiven Beurteilung des Klassifikators untersucht.

Wahrscheinlichkeitstabelle

Wird ein binärer Klassifikator auf ein Spektrum angewendet, so sind vier Versuchsausgänge möglich: Das Spektrum kann aus der Klasse *1* oder *2* stammen, und die Klassifikatorantwort kann ja (d. h. Zuordnung zu Klasse *1*) oder nein (d. h. Zuordnung zu Klasse *2*) sein (Tab. 1). Die Versuchsausgänge $(1, j)$ und $(2, n)$ sind richtig, $(1, n)$ und $(2, j)$ falsch.

Tabelle 1. *Die vier Versuchsausgänge bei Anwendung eines binären Klassifikators*

		Spektrenklasse 1	Spektrenklasse 2
Klassifikatorantwort	ja	$(1, j)$	$(2, j)$
	nein	$(1, n)$	$(2, n)$

Durch Klassifizierung einer Stichprobe bekannter Spektren erhält man Schätzwerte für die Wahrscheinlichkeiten $p(i, k)$ der vier Versuchsausgänge ($i = 1, 2$; $k = j, n$). $p(1)$ und $p(2)$ sind die Wahrscheinlichkeiten der Klassen *1* und *2* in der verwendeten Stichprobe. $p(j)$ und $p(n)$ sind die Wahrscheinlichkeiten, daß der Klassifikator bei dieser Stichprobe mit ja bzw. nein antwortet. Tab. 2 zeigt den Zusammen-

hang dieser Wahrscheinlichkeiten. Alle weiteren Beurteilungskriterien für Klassifikatoren werden aus den Werten der Tab. 2 abgeleitet und sind als Übersicht in Abb. 1 zusammengestellt.

Tabelle 2. *Wahrscheinlichkeitstabelle für einen binären Klassifikator*

$p(1,j)$	$+$	$p(2,j)$	$= p(j)$
$+$		$+$	$+$
$p(1,n)$	$+$	$p(2,n)$	$= p(n)$
$=$		$=$	$=$
$p(1)$	$+$	$p(2)$	$= 1$

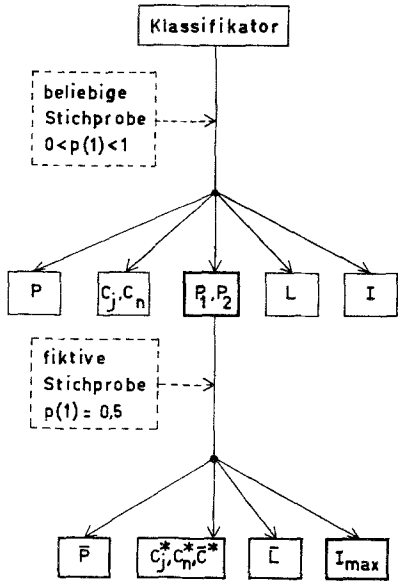

Abb. 1. Beurteilungskriterien für einen binären Klassifikator (siehe Text)

Stichprobe mit beliebiger Klassenaufteilung

Aus den Werten der Wahrscheinlichkeitstabelle (Tab. 2) lassen sich folgende Gütekriterien des Klassifikators direkt ableiten (Abb. 1):

a) Klassifizierungsfähigkeit für beide Klassen (P_1, P_2)

$$P_1 = \frac{p(1,j)}{p(1)} \qquad P_2 = \frac{p(2,n)}{p(2)} \tag{1}$$

P_1 und P_2 sind die bedingten Wahrscheinlichkeiten $p(j \mid 1)$ und $p(n \mid 2)$; sie geben an, wie groß die Wahrscheinlichkeiten sind, daß Spektren der Klasse *1* bzw. *2* richtig klassifiziert werden.

*b) Gesamtklassifizierungsfähigkeit (P)**

$$P = p\,(1) \cdot P_1 + p\,(2) \cdot P_2 = p\,(1,j) + p\,(2,n). \tag{2}$$

In der Literatur wurden Spektreninterpretationsverfahren bisher fast immer nur durch P charakterisiert. P ist der mittlere Prozentsatz von richtigen Klassifikatorantworten, wenn der Klassifikator auf Spektren angewendet wird, deren Klassenaufteilung durch $p\,(1)$ und $p\,(2)$ gegeben ist. Für unbekannte Spektren ist aber gerade diese Klassenaufteilung nicht gegeben.

c) Verläßlichkeiten der Antworten (C_j, C_n)

$$C_j = \frac{p\,(1,j)}{p\,(j)} \qquad C_n = \frac{p\,(2,n)}{p\,(n)} \tag{3}$$

C_j und C_n sind die bedingten Wahrscheinlichkeiten $p\,(1 \mid j)$ und $p\,(2 \mid n)$. C_j und C_n geben an, wie groß die Wahrscheinlichkeiten sind, daß die Antworten ja bzw. nein richtig sind[5, 6] — aber unter der Voraussetzung, daß die Klassenaufteilung der Stichprobe durch $p\,(1)$ und $p\,(2)$ gegeben ist**. C_j und C_n sollten eher als a-posteriori-Wahrscheinlichkeiten der Klassenzugehörigkeit nach der Klassifizierung aufgefaßt werden[4]. Die Wahrscheinlichkeit C, daß eine beliebige Antwort richtig ist, entspricht der Gesamtklassifizierungsfähigkeit P.

$$C = p\,(j) \cdot C_j + p\,(n) \cdot C_n = p\,(1,j) + p\,(2,n) = P. \tag{4}$$

d) Kosten der Klassifizierung (L)

Ordnet man jedem der vier Versuchsausgänge bestimmte Kosten $l\,(i,k)$ zu[7], so erhält man***

$$L = \sum_{i=1,2} \sum_{k=j,n} p\,(i,k) \cdot l\,(i,k) \tag{5}$$

Die Angabe der Kosten erlaubt es, beispielsweise die beiden falschen Versuchsausgänge $(1, n)$ und $(2, j)$ verschieden zu bewerten.

Werden die Kosten einer richtigen Entscheidung mit 0 und die einer falschen Entscheidung mit 1 festgesetzt, so erhält man

$$L = p\,(2,j) + p\,(1,n) = 1 - P. \tag{6}$$

---

* Die Bezeichnung P stammt von engl. predictive ability.
** Die Bezeichnung C stammt von engl. confidence.
*** Die Bezeichnung L stammt von engl. loss.

e) Information (I)

Die Informationstheorie liefert die im Versuch erhaltene Information (in bit) nach der Formel[4]

$$I = \sum_{i=1,2} \sum_{k=j,n} p(i,k) \cdot {}_2\log \frac{p(i,k)}{p(i) \cdot p(k)} \tag{7}$$

I gilt als objektive Charakteristik einer Prognose[8].

Von den bisher abgeleiteten Gütekriterien sind P, C_k, L und I abhängig von der Klassenaufteilung der Stichprobe und daher nicht geeignet, einen Klassifikator objektiv zu beurteilen. P, C_k, L oder I sind nur dann zum Vergleich von Klassifikatoren geeignet, wenn die Klassenaufteilungen der verwendeten Stichproben gleich sind.

Nur die gemeinsame Angabe der Werte für P_1 und P_2 ist unabhängig von der Klassenaufteilung der Stichprobe und daher geeignet, einen Klassifikator zu charakterisieren; nachteilig ist jedoch, daß zwei Zahlen angegeben werden müssen und daher der Vergleich von Klassifikatoren erschwert wird.

Stichprobe mit gleich wahrscheinlichen Klassen

Wird das Spektrum einer unbekannten Substanz klassifiziert, so nimmt man im allgemeinen an, daß es aus einer Stichprobe stammt, für die $p(1) = p(2) = 0{,}5$ gilt; d. h. Klasse *1* und *2* gleich wahrscheinlich sind. Für den Test eines Klassifikators steht eine derartig zusammengesetzte Stichprobe bekannter Spektren meist nicht zur Verfügung: einerseits gibt es in Spektrensammlungen meist nur einen geringen Prozentsatz von Substanzen, die eine bestimmte Partialstruktur ent-

Tabelle 3. *Wahrscheinlichkeitstabelle für einen binären Klassifikator, der durch P_1 und P_2 gegeben ist, bei Anwendung auf eine Stichprobe mit $p(1) = 0{,}5$*

$0{,}5\ P_1$	$0{,}5\ (1 - P_2)$	$0{,}5\ (1 + P_1 - P_2)$
$0{,}5\ (1 - P_1)$	$0{,}5\ P_2$	$0{,}5\ (1 - P_1 + P_2)$
0,5	0,5	1

halten, andererseits soll die Stichprobe möglichst groß sein. Hat man jedoch mit einer Stichprobe beliebiger Zusammensetzung $[0 < p(1) < 1]$ die Klassifizierungsfähigkeiten P_1 und P_2 für beide Klassen bestimmt, dann kann für eine fiktive Stichprobe mit $p(1) = 0{,}5$ eine neue Wahrscheinlichkeitstabelle errechnet werden (Tab. 3).

Aus den Werten der Tab. 3 lassen sich folgende Gütekriterien ableiten (Abb. 1):

a) Mittlere Klassifizierungsfähigkeit ($\bar{P}$)

$$\bar{P} = \frac{P_1 + P_2}{2} \tag{2}$$

$\bar{P}$ hat den Nachteil, daß Klassifikatoren mit unterschiedlichen Werten für P_1 und P_2 den gleichen Mittelwert liefern können.

b) Verläßlichkeiten der Antworten bei gleich häufigen Klassen (C_j^*, C_n^*)

$$C_j^* = \frac{P_1}{1 + P_1 - P_2} \qquad C_n^* = \frac{P_2}{1 - P_1 + P_2} \tag{9}$$

Daraus erhält man eine mittlere Verläßlichkeit[6] $\bar{C}^*$

$$\bar{C}^* = \frac{C_j^* + C_n^*}{2} \tag{10}$$

Für $\bar{C}^*$ gilt die gleiche Einschränkung wie für $\bar{P}$.

c) Mittlere Kosten ($\bar{L}$)

$$\bar{L} = 1 - \bar{P}. \tag{11}$$

$\bar{L}$ kann alternativ zu $\bar{P}$ oder $\bar{C}^*$ verwendet werden.

d) Maximale Information ($I_{\max}$)

$I_{\max}$ ergibt sich durch Einsetzen der Werte aus Tab. 3 in Formel (7). $I_{\max}$ ist die Information, die man erhält, wenn der Klassifikator auf eine Stichprobe mit $p\,(I) = 0{,}5$ angewendet wird — sie ist die maximal mit dem Klassifikator erzielbare Information. Aus Abb. 2 läßt sich für bekanntes P_1 und P_2 der Wert für $I_{\max}$ ablesen. Zwei Klassifikatoren haben auch dann verschiedene Werte von $I_{\max}$, wenn sie sich in $\bar{P}$ nicht unterscheiden (außer, wenn P_1 und P_2 bloß vertauscht sind). Wie bereits früher gezeigt wurde[4], liefern Klassifikatoren mit $P_1 + P_2 = 1$ keine Information; bei $P_1 + P_2 < 1$ sind die Antworten vertauscht. Aus Abb. 2 erkennt man, daß ein Klassifikator A mit $P_1 = 0{,}7$ und $P_2 = 0{,}9$ ($\bar{P} = 0{,}8$) mehr Information liefert ($I_{\max} = 0{,}296$ bit) als ein Klassifikator B mit $P_1 = P_2 = \bar{P} = 0{,}8$ ($I_{\max} = 0{,}232$ bit).

Ergebnis

Automatische Spektreninterpretationsverfahren sollten mit einer möglichst großen Stichprobe bekannter Spektren geprüft werden. Das Gütekriterium des Interpretationsverfahrens muß jedoch unabhängig von der Klassenaufteilung der Stichprobe sein oder sich auf eine Stichprobe beziehen, in der beide Klassen gleich häufig sind. Bei Benützung

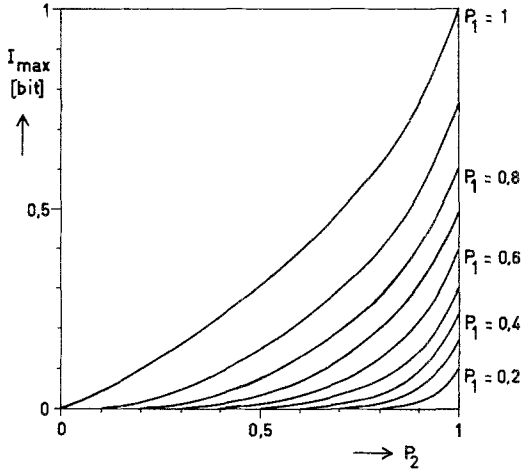

Abb. 2. Maximale Information I_{max} eines binären Klassifikators als Funktion der Klassifizierungsfähigkeiten P_1 und P_2 für die Klassen *1* und *2* (die beiden Klassen sind in der Stichprobe gleich häufig)

der angegebenen Formeln ist es nicht notwendig, eine — meist zu kleine — Stichprobe mit gleich häufigen Klassen zu verwenden.

Testet man ein Spektreninterpretationsverfahren mit einer Stichprobe bekannter Spektren und beliebiger Klassenaufteilung, so erhält man die beiden Klassifizierungsfähigkeiten P_1 und P_2 für Spektren beider Klassen. Die gemeinsame Angabe von P_1 und P_2 ist unabhängig von der Klassenaufteilung der verwendeten Stichprobe und daher zur mathematischen Charakterisierung eines Klassifikators geeignet.

Ein anderes Gütekriterium ist die maximale Information I_{max}, die man erhält, wenn man den Klassifikator (der durch P_1 und P_2 gegeben ist) auf eine (fiktive) Spektrenstichprobe anwendet, in der beide Klassen gleich häufig sind. I_{max} charakterisiert einen Klassifikator in einer einzigen Zahl.

Anschaulichere, aber weniger gut geeignete Gütekriterien sind die mittlere Klassifizierungsfähigkeit $\bar{P}$, die mittlere Verläßlichkeit der Antworten $\bar{C}^*$ oder die mittleren Kosten $\bar{L}$.

In Tab. 4 sind die in dieser Arbeit angeführten Gütekriterien für 3 verschiedene Klassifikatoren zusammengestellt. Die Werte für I_{max} zeigen, daß Klassifikator C besser ist als A und A besser ist als B. In diesem speziellen Fall liefern die Werte für $\overline{P}$ oder $\overline{C}^*$ die gleiche Reihenfolge. P liefert eine falsche Reihenfolge. Beim Vergleich der Werte für P_1 und P_2 kann man nicht entscheiden, welcher Klassifikator besser

Tabelle 4. *Vergleich von Klassifikatoren zur Interpretation niedrig aufgelöster Massenspektren. Klasse 1: Substanzen mit Stickstoff im Molekül, Klasse 2: Substanzen ohne Stickstoff im Molekül. Klassifikator A: „Entscheidungsebene"*[5], *B: „Nachbarspektrum aus einer Bibliothek"*[9], *C: „Entscheidungsebene mit toter Zone"*[10]. *N ist die Spektrenanzahl der Stichprobe. Alle Werte sind auf 2 Stellen gerundet*

Klassifikator:	A	B	C
N	250	500	330
$p\,(1)$	0,24	0,25	0,62
$p\,(2)$	0,76	0,75	0,38
P_1	0,83	0,74	0,94
P_2	0,93	0,94	0,83
P	0,90	0,89	0,90
L	0,10	0,11	0,10
I	0,38	0,33	0,48
C_j^*	0,92	0,93	0,85
C_n^*	0,85	0,78	0,93
$\overline{C}^*$	0,88	0,86	0,89
$\overline{P}$	0,88	0,84	0,89
$\overline{L}$	0,12	0,16	0,12
I_{max}	0,48	0,40	0,50

ist. In besonderen Fällen kann es allerdings darauf ankommen, daß eine der beiden Klassen möglichst gut erkannt wird; dann muß das entsprechende P_i zur Klassifikatorbeurteilung verwendet werden.

Für neu entwickelte Spektreninterpretationsverfahren sollten zumindest die Werte für P_1 und P_2 angegeben werden — nur dann können die Methoden verschiedener Autoren verglichen werden. Die bisher[1] meist übliche Angabe der Gesamtklassifizierungsfähigkeit P (predictive ability) erlaubt keine objektive Beurteilung.

Dank

Herrn Prof. Dr. *A. Maschka* danken wir für seine freundliche Unterstützung dieser Arbeit.

Literatur

[1] *P. C. Jurs* und *T. L. Isenhour*, Chemical Applications of Pattern Recognition. New York: Wiley. 1975.
[2] *B. R. Kowalski* und *C. F. Bender*, J. Amer. Chem. Soc. **94**, 5632 (1972).
[3] *K. Varmuza*, Mh. Chem. **105**, 1 (1974).
[4] *H. Rotter* und *K. Varmuza*, Org. Mass Spectrom. **10**, 874 (1975).
[5] *K. Varmuza* und *P. Krenmayr*, Z. Anal. Chem. **266**, 274 (1973).
[6] *P. Kent* und *T. Gäumann*, Helv. chim. Acta **58**, 787 (1975).
[7] *G. Meyer-Brötz* und *J. Schürmann*, Methoden der automatischen Zeichenerkennung. München: Oldenbourg. 1970.
[8] *A. M. Jaglom* und *I. M. Jaglom*, Wahrscheinlichkeit und Information. Berlin: VEB Dt. Verl. d. Wiss. 1965.
[9] *K. Varmuza*, Z. Anal. Chem. **268**, 352 (1974).
[10] *L. E. Wangen*, Computerized analysis of spectroscopic and chemical data by pattern classification techniques, S. 86, Diss. Univ. Washington, USA. 1971. Diss. Abstr. **32**, Nr. 72—7428 (1971).

Korrespondenz und Sonderdrucke:

Dr. K. Varmuza
Institut für Allgemeine Chemie
Technische Universität Wien
Lehárgasse 4
A-1060 Wien
Österreich