Chemometrie in der analytischen Chemie[1]

5. Diskriminanzanalyse in der Qualitätskontrolle

Eckhard Reh*
* Korrespondenz: Prof. Dr. E. Reh, Technische Hochschule Bingen, Email: e.reh@th-bingen.de

Einleitung

Die Diskriminanzanalyse ist eine Klassifizierungs-Methode, die in unterschiedlichsten Bereichen häufig eingesetzt wird. Ziel ist, Objekte an Hand ihrer unterschiedlichen Merkmale (Variablen) vorgegebenen Gruppen (Cluster) zuzuordnen. Zum Beispiel wird im Finanzwesen ein Kunde als kreditwürdig oder nicht kreditwürdig eingestuft, je nach seinen Merkmalen wie Schufa-Score, Einkommen, Zahl laufender Kredite etc..
Auch in der Chemie ist die Klassifizierung weit verbreitet, Merkmale sind hier Signale oder Konzentrationen vorliegender Proben, z.B. Nachweis gefälschten Biodiesels mittels NMR-Spektren [1].
Oftmals ist jedoch die Anwendung nicht optimal, teilweise fehlerhaft, zum Einen da eine unpassende Klassifizierungs-Methode eingesetzt wurde, daneben werden die besonderen Aspekte der Diskriminanzanalyse nicht erkannt bzw. genutzt. So wird immer wieder die lineare Diskriminanzanalyse mittels Diskriminanzfunktion verwendet, was auf Grund mehrerer Aspekte deplatziert ist (entschuldigend kann dies vielleicht damit erklärt werden, dass in der verwendeten Software meist keine Alternativen implementiert sind).
Es sollen daher zuerst die numerischen Grundlagen der Diskriminanzanalyse kurz behandelt werden (eine detaillierte theoretische Abhandlung siehe [2]) bevor sie mit Hilfe des Programms *CLUSTER* (Institut Chemometrie [3]) in zwei Anwendungsbeispielen eingesetzt wird.

Grundlagen

a) Vergleich diverser Klassifizierungs-Methoden

Im Unterschied zur Clusteranalyse ohne Vorinformationen liegen bei der Klassifizierung die Gruppierungen (Cluster) bereits vor mit sicher zugewiesenen Objekten (= Modellobjekte). Beispielsweise liegen in der Klinischen Chemie viele Blutproben vor mit unterschiedlichen, bekannten Blutparametern z.B. von Karzinom-Patienten (Cluster 1) und gesunden Probanden (Cluster 2). Ziel der Diskriminanzanalyse ist, neue Blutproben (= Probeobjekte) an Hand ihrer klinischen Parameter einem der beiden Clustern zuzuordnen, um evtl. entsprechend der Diagnose eine angemessene Therapie einzuleiten.

An Hand des Beispiels werden mehrere Aspekte für die Diskriminanzanalyse abgeleitet:
- es liegt ein 2-Cluster-Fall vor (kranke / gesunde Probanden)
  Oftmals liegen mehr Cluster vor, z.B. Qualitätskontrolle diverser Fruchtsäfte durch Kapillar-elektrophorese [4]. Die Diskriminanzanalyse sollte nur im 2-Clusterfall eingesetzt werden. Für mehrere Cluster kann sie prinzipiell mehrfach für die diversen Clusterpaare angewendet werden, es gibt aber alternative, z.T. bessere Methoden (z.B. kNN-, SIMCA-Methode).
- es liegt eine scharfe Zuordnung vor (ein Objekt kann nur einem Cluster angehören, einen 75% kranken Patienten gibt es nicht). Dies ist in der Diskriminanzanalyse unabdingbar.
  In der chemischen Anwendung muss dies nicht zwingend gegeben sein, so kann es sich bei einer chemischen Substanz um eine Säure oder ein Amin handeln, es gibt aber auch Verbindungen, die beiden Clustern zuzuordnen sind (z.B. Aminosäuren).
  In solchen Fällen (unscharfe Zuordnung) ist zwingend die SIMCA-Methode zu verwenden.
- ein Objekt muss immer einem vorgegebenen Cluster zugeordnet sein (Objekte die zu keinem Cluster passen, Ausreißer, gibt es nicht. Der Patient ist gesund oder krank).
  In vielen chemischen Anwendungen können aber Ausreißer auftreten, dann ist die kNN- als auch die SIMCA-Methode die richtige Wahl.

Die Situation wird unübersichtlicher dadurch, dass es diverse Varianten der Diskriminanzanalyse gibt, z.B. die euklidische, lineare, quadratische, regularisierte, PLS- oder Maximum-Likelihood-Diskriminanzanalyse.

1 Reh, E., GIT, 6, 13-17, 2019

b) Vergleich euklidische, quadratische Diskriminanzanalyse

Abbildung 1 skizziert den Einsatz von euklidischer bzw. quadratischer Diskriminanzanalyse in einem einfachen Fall mit 2 Merkmalen (Variablen x_1, x_2).

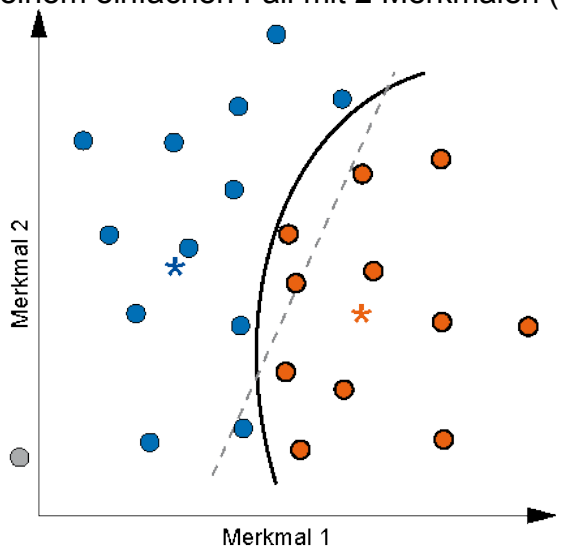


Abb. 1: 2-Cluster-Klassifizierung
- euklidische Diskriminanzanalyse (Klassifikator: graue Linie, gestrichelt),
- quadratische Diskriminanzanalyse (Klassifikator: schwarze Kurve),

grauer Punkt: zu klassifizierendes Objekt; *: Zentroid des jeweiligen Clusters

Im Fall der oft verwendeten euklidischen Diskriminanzanalyse ist der Klassifikator („Grenze“) zwischen beiden Clustern eine Gerade bzw. Hyperebene im n-dimensionalen Raum (Abbildung 1, gestrichelte, graue Linie). Es ist offensichtlich, dass dies, wie bei den meisten anderen Varianten, in kritischen Fällen unpassend ist. Nur die quadratische Diskriminanzanalyse liefert eine gekrümmte Kurve / Hyperfläche als Klassifikator und damit nahezu immer bessere Resultate (Abbildung 1, schwarze Kurve).

Darüber hinaus gibt es zwei numerische Ansätze zur Bestimmung des Klassifikators, oft wird die aufwendige Berechnung einer Diskriminanzfunktion verwendet (numerische Details vgl. [2]). Sowohl bei euklidischer als auch quadratischer Diskriminanzanalyse ist der Klassifikator jedoch einfach definiert durch Punkte gleicher Distanz zu Cluster 1 bzw. 2.

Der direkte Ansatz zur Klassifizierung eines neuen Objekts basiert auf der Verwendung der Distanz des Objekts zu Cluster 1 bzw. Cluster 2, das Objekt wird dem Cluster zugeordnet, zu dem (bzw. zu dessen Schwerpunkt, Zentroid, vgl. Abbildung 1) es die geringere Distanz hat (der Klassifikator ist hier zur Klassifizierung nicht explizit nötig).

- Für den Fall der euklidischen Distanz eines Objekts zum Cluster 1, d(O, C1), gilt

$$d(O,C1)=\sqrt{(\boldsymbol{x}_O-\boldsymbol{z}_1)^T(\boldsymbol{x}_O-\boldsymbol{z}_1)} \qquad (1)$$

mit $\mathbf{x_O}$: Merkmals-Vektor Objekt O; $\mathbf{z}_1$: Zentroid-Vektor Cluster 1; T: transponierter (Zeilen-)Vektor

Gleichung 1 definiert den vektoriellen Abstand zwischen Objekt O und dem Zentroid des Clusters 1, analog für die Distanz zu Cluster 2.
- Für die Distanz zu Cluster 1 wird bei der quadratische Diskriminanzanalyse die Mahalanobis-Distanz verwendet:

$$d(O,C1)=\sqrt{(\boldsymbol{x}_O-\boldsymbol{z}_1)^T\boldsymbol{C}_1^{-1}(\boldsymbol{x}_O-\boldsymbol{z}_1)} \qquad (2)$$ analog für Cluster 2

mit $\boldsymbol{C_1^{-1}}$: inverse Varianz-Kovarianz-Matrix berechnet aus Varianz-Kovarianz-Matrix der Merkmale von Cluster 1 (siehe Anhang)

Die Bestimmungsgleichung für den Klassifikator ist in beiden Fällen definiert durch Punkte X_P mit gleicher Distanz zu Cluster 1 und 2, es gilt: $d(X_P, C1) = d(X_P, C2)$ (3)
Nur bei der quadratischen Diskriminanzanalyse verbleiben quadratische Terme, es folgt eine gekrümmte Kurve, Hyperfläche als Klassifikator (Quadrik).

Ein Vergleich euklidische, quadratische Diskriminanzanalyse für 3 entsprechende Merkmale ist im Anhang dargestellt.

Nachteilig bei der quadratischen Diskriminanzanalyse ist, dass die numerische Berechnung einer inversen Matrix ($\boldsymbol{C_1^{-1}}$, $\boldsymbol{C_2^{-1}}$) kritisch sein kann. Der Grund liegt zumeist in der Qualität der Daten, wenn z.B. viele Merkmale voneinander abhängig sind (Kollinearität) [6]).
Daher ist meist eine Optimierung der eingesetzten Merkmale sinnvoll.

c) Selektierung Merkmale
Während in vielen Publikationen unterschiedlichste Klassifizierungs-Methoden verglichen und bewertet werden, wird oftmals ein zentraler Aspekt einer optimalen Differenzierung außer Acht gelassen. Entscheidend ist, welche Merkmale (Konzentrationen, Signale) für den aktuellen Fall eingesetzt werden. Es ist in den meisten Fällen nicht hilfreich, möglichst viele Parameter zu verwenden, essentiell ist, die für die vorliegende Aufgabenstellung relevanten Merkmale zu benutzen.
Anders als bei der Clusteranalyse stehen in der Klassifizierung durch die bekannte Zugehörigkeit der Modell-Objekte zu den Clustern Methoden zur Verfügung, die relevanten Merkmale zu selektieren (feature selection).
Zur Beurteilung der Qualität der Klassifizierung wird als Maßzahl meist der Anteil korrekt zugeordneter Probe-Objekte, %CC, verwendet. Es gilt

$$\%CC = 100 \frac{1}{n} \sum_{c=1}^{m} TP_c \quad (4)$$

mit TP_c: Zahl korrekt zugeordneter Objekte in Cluster c; n: Gesamtzahl Objekte; m: Clusterzahl

Der %CC-Wert kann z.B. ermittelt werden, indem alle Modell-Objekte zusätzlich als Probe-Objekte neu zugeordnet werden (Autoprediction, suboptimal da Probe- = Modell-Objekte).

- Rang eines Merkmals
Ein Merkmal (z.B. Konzentration Ba) hat einen hohen Rang, wenn der Unterschied seines Mittelwerts in den Clustern groß ist. Im 2-Cluster-Fall kann hierzu die Student-t-Prüfgröße herangezogen werden, allgemein einsetzbar im n-Cluster-Fall ist der Fisher-Wert:

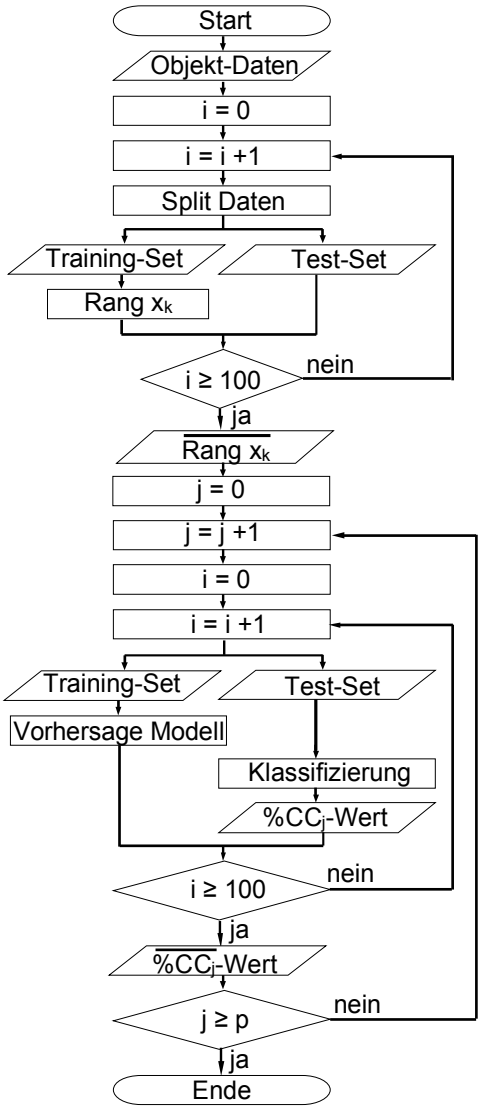
$$\hat{f}_k = \frac{\sum_{c=1}^{m} n_c (\bar{x}_{ck} - \bar{x}_k)^2}{\sum_{c=1}^{m} (n_c - 1) s_{ck}^2} \quad (5) \qquad \text{für Variable k}$$

mit $\bar{x}_{ck}$: Mittelwert k-te Variable in Cluster c; $\bar{x}_k$: Mittelwert k-te Variable über alle Cluster;
n_c: Probezahl in Cluster c; m: Clusterzahl; s_{ck}: Standardabweichung k-te Variable in Cluster c

Der Fisher-Wert liefert, nach fallender Größe sortiert, den Rang der Merkmale.

Zusätzlich kann mit Hilfe von Monte-Carlo-Methoden die Signifikanz der Merkmale bestimmt werden [5]. Dies ist jedoch für die Ermittlung der relevanten Merkmale nicht zwingend erforderlich.

- Relevante Merkmale
Nur Merkmale, die auch einen wesentlichen Beitrag leisten, sollten zur Klassifizierung herangezogen werden. Die numerische Behandlung zur Ermittlung solcher, relevanter Merkmale (Bootstrapping) ist aufwändiger und soll hier phänomenologisch behandelt werden (detaillierte Diskussion siehe [5]).



Die Modellobjekte werden zufallsbedingt in Trainings- und Test-Set (2/3 und 1/3) aufgeteilt, basierend auf den Trainings-Objekten werden die Test-Objekte klassifiziert d.h. einem der beiden Cluster zugeordnet (auf Grund der Zufallseinteilung wird dieses 100 x wiederholt).

In einem 1. Teil wird aus den jeweiligen Trainings-Objekten der Rang jedes Merkmals x_k bestimmt, bzw. aus den Wiederholungen der Mittelwert jedes Rangs.

In einem 2. Teil wird danach die Klassifizierung mit einem Merkmal (mit dem höchsten Rang) durchgeführt und die Güte $\overline{\%CC_1}$ ermittelt.
Im nächsten Zyklus wird das 2. Merkmal (mit nächst geringerem Rang) hinzu genommen, usw. jeweils 100 x (Bootstrapping mit Mittelwertbildung).
Wird die Güte der Klassifizierung ($\overline{\%CC_j}$) durch ein weiteres Merkmal j nicht mehr verbessert, sind die verbliebenen Merkmale (mit geringem Rang) nicht relevant und sollten unberücksichtigt bleiben.
Abbildung 2 gibt das Procedere als Fließschema wieder.

Abb. 2: Fließschema Bootstrapping relevante Merkmale
(p: Merkmalzahl)

Beispiel 1: Keramik-Proben

Aufgabenstellung ist die Bewertung von Keramik-Proben [7] und Unterscheidung in Ton- bzw. C-haltiges Material mit Hilfe von Element-Konzentrationen (Ti, Sr, Ba, Mn, Cr, Ca, Al, Fe, Mg, Na, K).

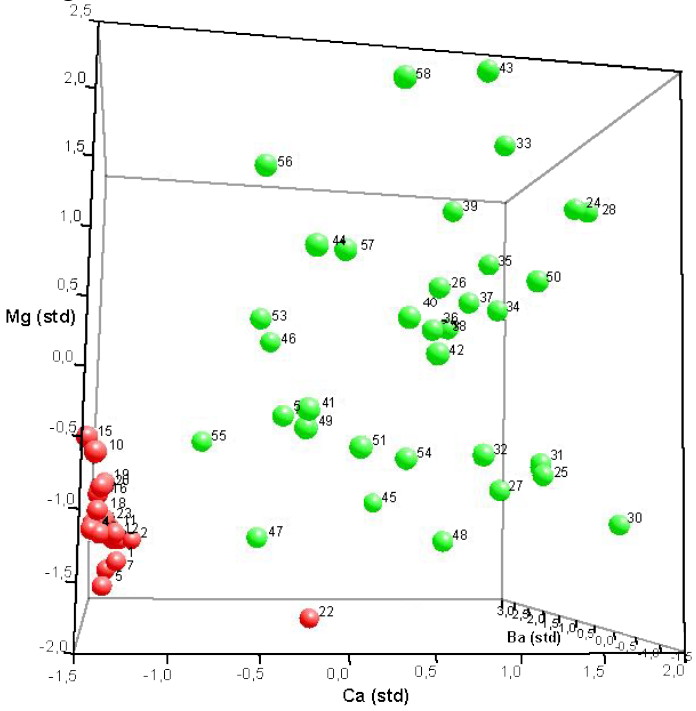


Abbildung 3 zeigt den Clusterplot,
die Lokalisierung der Proben im 3D-Raum von je 3 Variablen, Konzentrationen Ba, Ca, Mg (standardisiert, siehe Anhang).

Abb. 3: Clusterplot Keramik-Objekte,
Merkmal Ba, Ca, Mg (standardisiert)
rot: Cluster mit Ton-haltigen Objekten)

Tabelle 1 zeigt einen Auszug des Validierungs-Report des verwendeten Programms *CLUSTER* für die Keramik-Objekte und listet u.a. den Rang der Merkmale auf, hohen Rang haben Ca, Ba, Mg (nicht signifikant sind Cr und K, Signfikanz-Werte nicht angegeben).

	Ti	Sr	Ba	Mn	Cr	Ca	Al	Fe	Mg	Na	K
Fisher-Wert	0,14	0,48	1,52	0,27	0,06	2,55	0,76	0,17	0,86	0,12	0,06
Rang	8	5	2	6	11	1	4	7	3	9	10

Diskriminanz-Analyse euklidisch

%CC-Wert	94,83	(Autoprediction)

Diskriminanz-Analyse quadratisch

%CC-Wert	100,00	(Autoprediction)

Kontingenz-Tabelle:

	Objekte Cluster 1	Objekte Cluster 2
Cluster 1 zugeordnet	23	3
Cluster 2 zugeordnet	0	32

	Objekte Cluster 1	Objekte Cluster 2
Cluster 1 zugeordnet	23	0
Cluster 2 zugeordnet	0	35

Erhöhung Merkmalzahl mit fallendem Rang (Bootstrapping):

Zugefügtes Merkmal	Zahl Merkmale j	$\overline{\%CC_j}$
Ca	1	91,20
Ba	2	96,47
Mg	3	94,27
Al	4	95,80
Sr	5	95,53
Mn	6	95,20
Fe	7	94,93
Ti	8	94,73
Na	9	94,87
K	10	95,20
Cr	11	95,27

Zugefügtes Merkmal	Zahl Merkmale j	$\overline{\%CC_j}$
Ca	1	97,00
Ba	2	96,87
Mg	3	96,87
Al	4	96,60
Sr	5	95,60
Mn	6	93,60
Fe	7	93,93
Ti	8	93,20
Na	9	92,80
K	10	90,73
Cr	11	88,93

Tab. 1: Auszug Validierungs-Resultate *CLUSTER*

Die Kontingenz-Tabelle der euklidischen Diskriminanzanalyse zeigt, dass von 23 Objekten des Cluster 1 alle korrekt Cluster 1 zugeordnet wurden, von den 35 Objekten des Cluster 2 aber 3 falsch dem Cluster 1 zugeordnet wurden, der %CC-Wert ist entsprechend nur 94,83 % (Autoprediction). Bei der quadratischen Diskriminanzanalyse wurden alle Objekte korrekt den beiden Clustern zugeordnet, der %CC-Wert ist 100 % (Autoprediction).
Der mittlere $\overline{\%CC_j}$-Wert (Bootstrapping) steigt bei Erhöhung auf 2 Merkmale (mit Rang 1 und 2), danach fällt der Anteil der korrekt zugeordneten Objekte ab, d.h. weitere Merkmale verschlechtern die Zuordnung z.B. auf Grund der geringeren Bedeutung für die Differenzierung bzw. höherem Rauschen.
Die Konsequenz ist, das statt einer ICP-MS-Analyse vieler Elemente eine einfache Atom-Emissions-Messung weniger Erdalkali-Elemente ausreichen (Ca, Ba, evtl. Mg).

Beispiel 2: Lymph-Gewebe

Zur schnellen Differenzierung von Lymph-Proben (cancerogen, normal) wurde der wäßrige Gewebe-Extrakt mittels Raman-Spektroskopie untersucht [8]. Die Spektren wurden in feste, konsekutive Intervalle von je 50 cm^{-1} unterteilt und die mittlere Bandenhöhe in die Klassifizierung einbezogen (dies entspricht dem durch die FDA propagierten Vorgehen für kontinuierliche Messreihen).

Der Distanzplot Abbildung 4 deutet auf eine akzeptable Differenzierung beider Cluster hin.
Im Distanzplot wird die Distanz eines Objektes zum Zentroid der beiden Cluster aufgetragen, d.h. Objekt 48 hat eine kleine Distanz zu Cluster 1 (dem es angehört) und große Distanz zu Cluster 2.

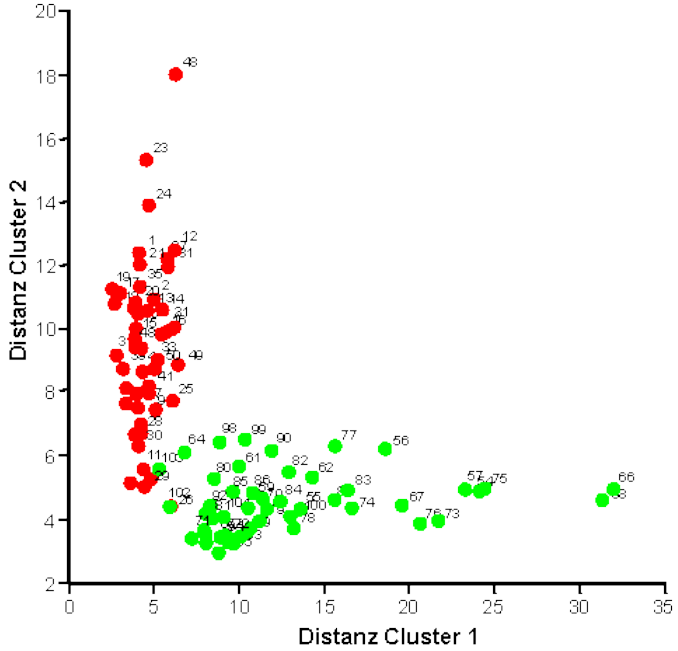


Abb. 4: Distanzplot Lymph-Objekte (Mahalanobis Distanz)

Die Optimierung und Validierung der Modell-Objekte zeigt Tabelle 2.
Bei 174 Merkmal-Paaren liegen hohe Korrelationen vor ($r_{x_j,x_k} > 0{,}9$, siehe Anhang, Werte nicht explizit angegeben), dies ist nicht verwunderlich beim Einsatz äquidistanter Raman-Spektralbereiche.
Die Fischer-Werte sind generell sehr klein, haben geringe Bedeutung zur Differenzierung (nur die 4 Merkmalsbereiche 925, 1225, 1425, 1625 $\pm$ 25 cm^{-1} sind signifikant, Signifikanz-Werte nicht angegeben).
Die Kontingenz-Tabelle (quadratische Diskriminanzanalyse, nicht explizit angegeben) zeigt, dass von 53 Objekten des Cluster 1 nur 52 korrekt Cluster 1 zugeordnet wurden, von den 50 Objekten des Cluster 2 wurde ebenfalls eins falsch dem Cluster 1 zugeordnet, der %CC-Wert ist ~ 98 % (Autoprediction).
Der mittlere $\overline{\%CC_j}$-Wert (Bootstrapping) steigt mit Erhöhung der Merkmal-Zahl bis zum 6. Merkmal gravierend, danach nur gering bis auf ~ 80 %.

* * * Optimierung Merkmale * * *

Bereich [cm^{-1}]	Fisher-Wert	Rang
675	0,001	21
725	0,002	20
775	0,002	19
825	0,017	11
875	0,014	12
925	0,038	4
975	0,030	5
1025	0,018	10
1075	0,008	15
1125	0,002	18
1175	0,009	14
1225	0,039	3
1275	0,027	6
1325	0,019	7
1375	0,018	9
1425	0,075	1
1475	0,000	22
1525	0,002	17
1575	0,007	16
1625	0,071	2
1675	0,018	8
1725	0,013	13

Erhöhung Merkmalzahl mit fallendem Rang (Bootstrapping):

Zugefügtes Merkmal	Zahl Merkmale j	$\overline{\%CC_j}$
1425	1	56,57
1625	2	56,26
1225	3	55,77
925	4	62,43
975	5	65,86
1275	6	72,49
1325	7	71,77
1675	8	71,49
1375	9	71,66
1025	10	72,57
825	11	73,23
875	12	73,69
1725	13	74,49
1175	14	75,40
1075	15	76,69
1575	16	77,97
1525	17	76,97
1125	18	77,97
775	19	79,26
725	20	78,29
675	21	78,29
1475	22	79,17

Tab. 2: Auszug Validierungs-Resultate *CLUSTER*

Abbildung 5 gibt im 3D-Clusterplot der 3 Merkmale mit höchstem Rang eine visuelle Abschätzung der kritischen Differenzierung (im höher dimensionalen Raum wird dies evtl. günstiger sein).

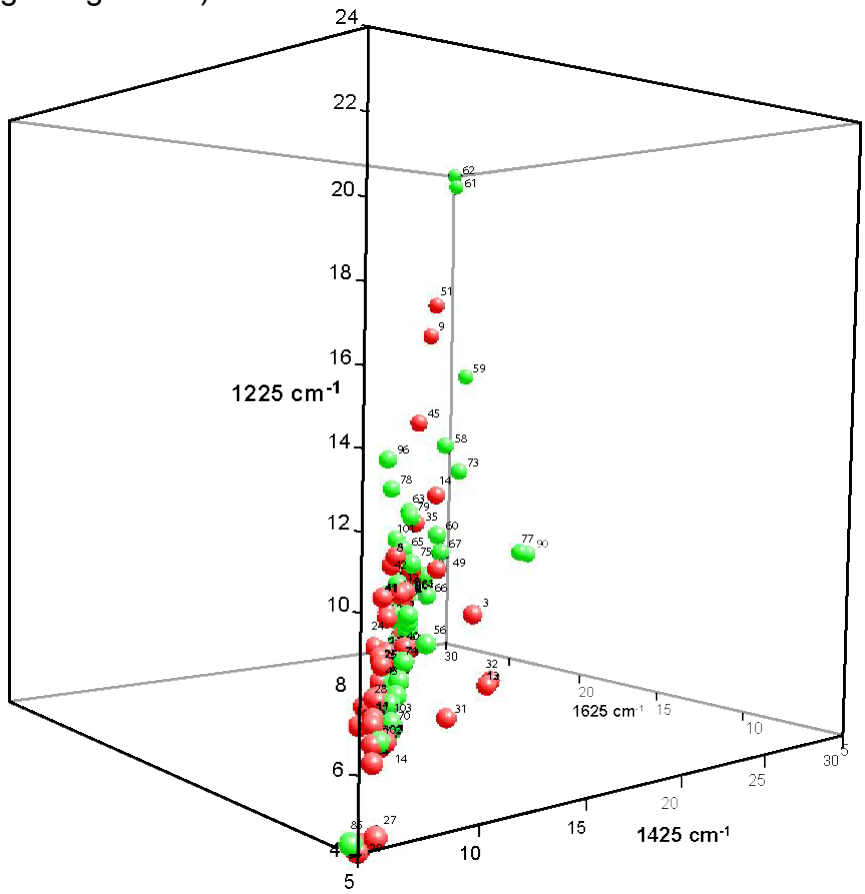


Abb. 5: Clusterplot Lymph-Proben, Merkmal 1225, 1425, 1626 ± 25 cm^{-1}

Zur Verbesserung der Klassifizierung sollte primär ein optimaler Ansatz zur Ermittlung diskreter Merkmale herangezogen werden, die Verwendung des gesamten spektralen Bereichs in Form fester Intervalle (FDA-Procedere) kann nicht empfohlen werden. Für die Extraktion von diskreten Merkmalen aus kontinuierlichen Messreihen (feature extraction) wie z.B. IR-Spektren oder ungetrennter GPC-Chromatogramme gibt es bessere Ansätze (z.B. successive projection, loading spectrum [9]).

Zusammenfassung:

Für die Diskriminanzanalyse muß festgehalten werden:
- sie ist nur im 2-Cluster-Fall bei scharfer Zuordnung ohne Ausreißer sinnvoll,
- in der Chemie ist die quadratische Diskriminanzanalyse zumeist die Methode der Wahl,
- die Bestimmung / Verwendung der relevanten Merkmale ist sehr wichtig.

Die quadratische Diskriminanzanalyse bietet sehr gute Klassifizierungen auch in kritischen Aufgabenstellungen bei optimaler Wahl relevanter Merkmale.
Die kNN- oder SIMCA-Methode sind valide Alternativen zur Diskriminanzanalyse.

Der Einsatz aufwendiger Algorithmen wie z.B. support vector machines, learning vector quantisation oder artificial neuronal networks ist in der Chemie zumeist nicht nötig, sondern birgt zudem die Gefahr einer nicht realen Überinterpretation (overfitting) [5].

Anhang

1. Varianz-Kovarianz-Matrix $\boldsymbol{C_1}$ (Objekte in Cluster 1 mit 2 Merkmalen x_1, x_2)

$$\boldsymbol{C_1} = \begin{pmatrix} var(\boldsymbol{x_1}) & cov(\boldsymbol{x_1}, \boldsymbol{x_2}) \\ cov(\boldsymbol{x_2}, \boldsymbol{x_1}) & var(\boldsymbol{x_2}) \end{pmatrix} \quad (6)$$

Varianz $var(\boldsymbol{x_k}) = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{ik} - \bar{x}_k)^2$ (7) Kovarianz $cov(\boldsymbol{x_j}, \boldsymbol{x_k}) = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ (8)

mit n_1: Objekt-Zahl in Cluster 1; Merkmal j und k

2. Standardisierung x^s_{ik} (bei sehr unterschiedlichen Merkmalgrößen)

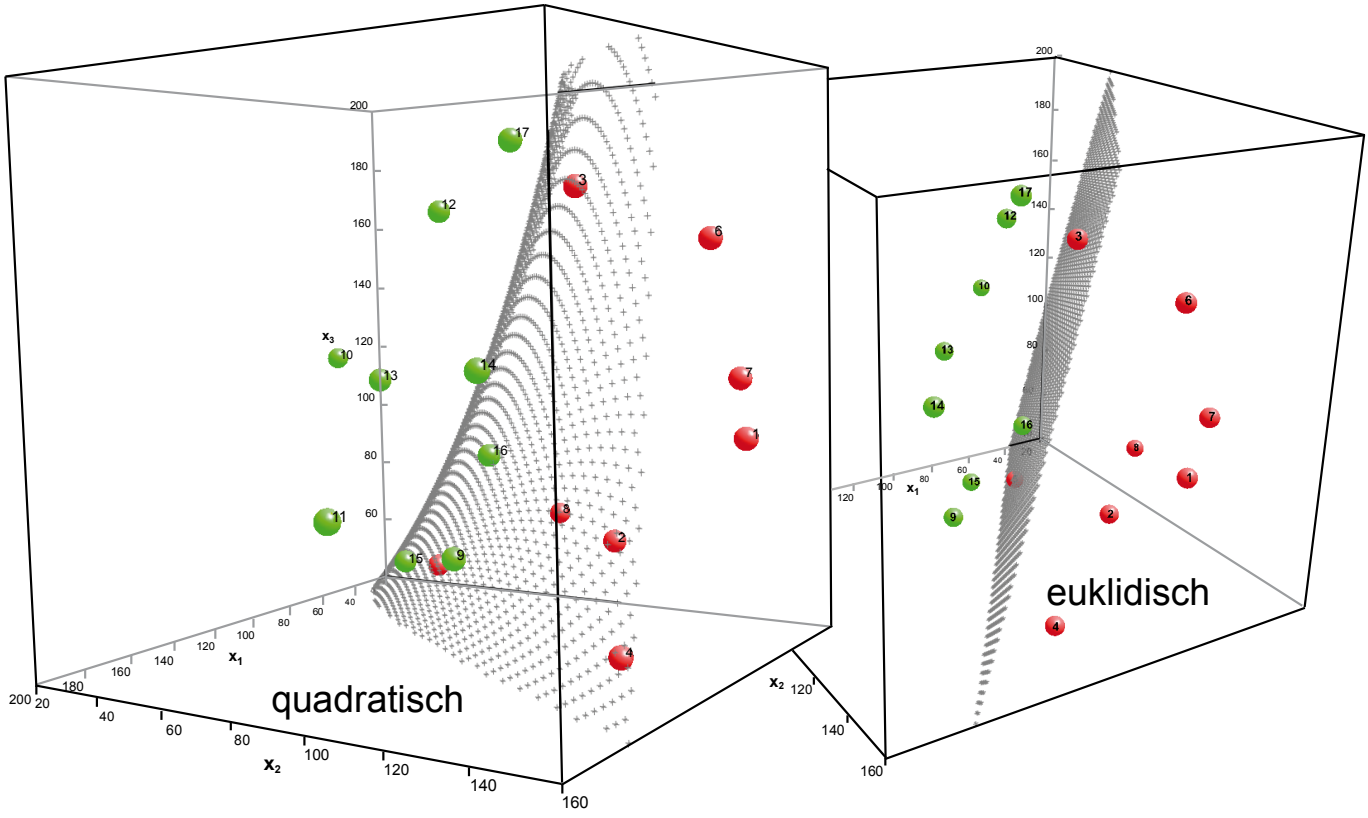
$$x^s_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k} \quad (9)$$

mit s_k: Standardabweichung Merkmal k (Freiheitsgrad n)

3. Korrelationskoeffizient $r_{x_j x_k}$ (für die Merkmale j und k aller Modellobjekte)

$$r_{x_j x_k} = \begin{cases} \frac{cov(\boldsymbol{x_j}, \boldsymbol{x_k})}{\sqrt{var(\boldsymbol{x_j})\, var(\boldsymbol{x_k})}} & \text{für } j \neq k \\ 1{,}0 & \text{für } j = k \end{cases} \quad (10)$$

4. Vergleich quadratische, euklidische Diskriminanzanalyse (3 Merkmale)



Literatur

[1] Álvaro C. Neto, et al., Quality control of ethanol fuel: Assessment of adulteration with methanol using ^{1}H-NMR, Fuel, 135, 387-392, 2014
[2] Reh, E., Diskriminanzanalyse, www.chemometrie.info/literatur-titel.html
[3] www.chemometrie.info/statistikcluster-zielsetzung.html
[4] María Navarro-Pascual-Ahuir et al, Quality control of fruit juices by using organic acids determined by capillary zone electrophoresis with poly(vinyl alcohol)-coated bubble cell capillaries, Food Chemistry, 18, 596-603, 2015
[5] Brereton, R.G., Chemometrics for Pattern Recognition, Wiley-VCH, Weinheim, 2009
[6] Reh, E. Chemometrie: Grundlagen der Statistik, Numerischen Mathematik und Software Anwendungen in der Chemie. Walter de Gruyter GmbH & Co KG, 2017
[7] Bruno, P., Caselli, M., Curri, M.L., Genga, A., Striccoli, R, Traini, A., Chemical characterisation of ancient pottery from south of Italy by Inductively Coupled Plasma Atomic Emission Spectroscopy (ICP-AES): Statistical multivariate analysis of data, Anal. Chim. Acta., 410, 193-202, 2000
[8] Lloyd, G.R., Orr, L.E., Christie-Brown, J., Analyst, 138, 3900-3908, 2013
[9] Reh, E., Feature Extraction with Loading Spectra in Cluster-Analysis and Classification, Talanta, in press 2019