Automatische Klassifikation

Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten

(Cluster-Analyse)

Von

Dr. rer. nat. Hans Hermann Bock

Technische Universität Hannover

Mit 54 Abbildungen

VANDENHOECK & RUPRECHT IN GÖTTINGEN

Studia Mathematica / Mathematische Lehrbücher

Herausgegeben von
Karl Peter Grotemeyer, Bielefeld
Dietrich Morgenstern, Hannover / Horst Tietz, Hannover

Band XXIV

ISBN 3-525-40130-2

Herstellung: Hubert & Co., Göttingen

Vorwort

Das vorliegende Buch behandelt Methoden und Verfahren, die unter dem Namen „automatische Klassifikation" oder „Cluster-Analyse" bekannt sind. Dabei geht es um das Problem, eine i. a. große Menge zunächst ungeordneter Objekte (z. B. Personen oder Dokumente) aufgrund von Ähnlichkeit und sachlicher Verwandtschaft in kleinere, homogene und praktisch nützliche Klassen oder Gruppen zu zerlegen. Eine solche Klassifikation von Objekten erweist sich in vielen Situationen als zweckmäßig oder notwendig. Sie erlaubt unter anderem, die Struktur der betrachteten Objektmenge in vereinfachter und übersichtlicher Weise darzustellen und hierdurch die zwischen den Gruppen bestehenden Zusammenhänge leicht zu erkennen. Auch können die einzelnen, homogenen Objektklassen häufig als „Objekt-Typen" mit spezifischen Eigenschaften interpretiert und dann – wie etwa bei Krankheitstypen – jeweils gesondert behandelt werden.
In der Vergangenheit ging man zur Lösung von Gruppierungsproblemen weitgehend empirisch und intuitiv vor; das Beispiel der bereits von LINNÉ (1758) eingeführten und heute noch gebräuchlichen Klassifikation der Lebewesen zeigt, daß auf diese Weise sehr nützliche Ergebnisse erzielt werden können. Andererseits ist offensichtlich, daß bei intuitivem Vorgehen die spezielle Struktur des jeweiligen Anwendungsgebiets sowie die Geschicklichkeit und Sachkenntnis des Untersuchenden eine vorrangige Rolle spielen und zu einer subjektiven Interpretation der Wirklichkeit führen können. Letzteres zeigt sich etwa dann, wenn das breite Spektrum politischer Meinungen in zwei Klassen „links" und „rechts" aufgeteilt wird, die je nach dem Standpunkt des Betrachters sehr unterschiedlich zusammengesetzt sein werden.

Die „automatische Klassifikation" verdankt ihre Entwicklung dem Wunsch, den Klassifikationsprozeß systematisch und quantitativ zu erfassen und durch Berücksichtigung numerischer Kriterien die Güte von Gruppierungen „objektiv" zu vergleichen. Aus diesem Bestreben wurden in den letzten Jahren zahlreiche Verfahren ersonnen, die zur Konstruktion optimaler, fast-optimaler oder „zweckmäßiger" Klassifikationen dienen und die hierbei ausschließlich mathematisch-statistische Hilfsmittel verwenden. Die Bevorzugung mathematischer Methoden ist auch dadurch bedingt, daß die Analyse der immer größer werdenden Datenmengen oft nur bei Benutzung automatischer Rechenanlagen möglich ist; deren Aufkommen hat umgekehrt die Entwicklung numerischer Klassifikationsverfahren sehr begünstigt.

Die vorliegende Monographie bietet eine Einführung in die Methoden und Modelle der automatischen Klassifikation und gibt einen Überblick über die praktisch wichtigen Verfahren. Im Gegensatz zu Autoren, die dieses Thema im Hinblick auf spezielle, z. B. biologische Anwendungen behandeln, wurde hier das Hauptgewicht auf eine präzise Darstellung der mathematischen Grundlagen gelegt und versucht, Struktur und Eigenschaften der betrachteten Klassifikationen exakt – soweit als möglich – zu charakterisieren. Das Buch ist in drei Teile gegliedert.

Nach einer Einführung, in der die behandelte Fragestellung anhand von Beispielen erläutert wird, werden in Teil I verschiedene Möglichkeiten aufgezeigt, um die Ähnlichkeit oder Unähnlichkeit von Objekten sowie die Homogenität von Objektmengen zu messen. Der Teil II behandelt disjunkte Gruppierungen (bei denen die Klassen sich nicht überschneiden dürfen) und untersucht zunächst ein statistisches Modell; dabei wird die Optimalität gewisser Klassifikationsverfahren bewiesen. Nach Diskussion verschiedener Kriterien zur Bewertung von Gruppierungen betrachten wir numerische Methoden, darunter Verfahren zur Analyse von Verteilungsmischungen sowie sequentielle Methoden. Es folgen graphentheoretische Modelle.

Teil III ist den nichtdisjunkten und hierarchischen Klassifikationen gewidmet. Wir beschreiben insbesondere die Konstruktion maximaler Cliquen und die von NEEDHAM/PARKER-RHODES entwickelten Methoden. Bei hierarchischen Klassifikationen wird der Zusammenhang zwischen Dendrogrammen und ultrametrischen Distanzmaßen sowie die Optimalität von Hierarchien diskutiert. Zur praktischen Konstruktion von Hierarchien dienen u. a. agglomerative und divisive Verfahren. Ein Abschnitt über nichtdisjunkte Hierarchien schließt den Band ab.

Zur Lektüre dieses Buches sind elementare Kenntnisse aus der Statistik und der Matrizenrechnung wünschenswert. Da die Thematik gerade den Praktiker interessieren dürfte, werden die grundlegenden Begriffe und Verfahren sehr ausführlich dargestellt und anschaulich kommentiert. Hierdurch ist gesichert, daß auch der mathematisch wenig Vorgebildete aus der Lektüre Nutzen zieht, sofern er die theoretischen Beweise übergeht; deren Ende ist jeweils durch das Zeichen ■ angedeutet. Eventuell ungewohnte Bezeichnungsweisen werden im Anhang erläutert. Abschnitte, die beim ersten Lesen übersprungen werden können oder weitgehend theoretisch orientiert sind, wurden mit einem Stern * versehen.

Zum besseren Verständnis der allgemeinen Methoden habe ich eine Reihe von Abbildungen und Zahlenbeispielen eingefügt; letztere besitzen ausschließlich

exemplarischen Charakter. Spezielle, aus der Praxis stammende Anwendungen werden nicht behandelt, da hierbei die Anwendbarkeit der betreffenden Klassifikationsmodelle ausführlich diskutiert werden müßte; stattdessen findet man zahlreiche Hinweise auf Arbeiten, in denen solche Anwendungen beschrieben sind.

An dieser Stelle möchte ich Herrn Prof. Dr. D. Morgenstern für sein förderndes Interesse an dieser Arbeit herzlich danken. Auch danke ich meiner Frau für die Geduld, mit der sie das Manuskript und dessen Überarbeitungen geschrieben hat. Mein Dank gilt außerdem den Herausgebern der Reihe STUDIA MATHEMATICA sowie dem Verlag für die sorgfältige Ausstattung dieses Bandes.

Ich hoffe, daß das Buch dazu verhilft, die Möglichkeiten und Verfahren automatischer Klassifikationsverfahren einem großen Kreis von Praktikern bekannt zu machen, und daß es zur Behandlung der zahlreichen, auf diesem Gebiet noch ungelösten Fragen theoretischer wie praktischer Art anregt.

Hannover, im April 1973 H. H. Bock