

Book Reviews

DIDAY, E. et al: *Optimisation en Classification Automatique*. (Optimisation in Numerical Taxonomy). (In French)

Le Chesnay, FR: Inst. National de Recherche en Informatique et en Automatique 1979. 2 vols. 896 p., 120ff. ISBN 20-7261-0219-0

This is a book on classification and clustering methods and their use in exploratory data analysis. The book is mathematically oriented and has little overlap with other monographs on cluster analysis. Consider a set E of objects i , each one described by some observation or data vector x_i . The problem is to subdivide this set E into a certain number of homogeneous groups or classes, i.e. such that the elements of each class will be as 'similar' as possible (and near to some suitable class representative) whilst different classes are most 'dissimilar' in some sense. Hereby 'similarity' resp. 'dissimilarity' of objects will be measured by some distance of the corresponding feature vectors. It is evident that the resulting classification of objects may be useful for information retrieval or documentation sciences, for organisational purposes in economics, for pattern recognition, psychology a.s.o. In particular, the resulting groups may be interpreted as 'types' or can serve for information (data) reduction.

Diday et al describe a lot of cluster analytic methods which were developed by a French research group at INRIA and which all look for those classifications (partitions) of E which are (nearly) optimal according to some given optimality criterion W . Most algorithms are modifications of the well-known k -means method and operate by iteratively determining a minimum-distance partition and correspondingly adapting the former class representatives ('dynamic clusters method'). The principle and the convergence of this method are established in § 1. The following 24 chapters show – this is the true value of this book – that by suitably defining the notions 'distance' and 'class representative' the method may be used for many practical situations or optimality criteria: The usual quadratic distances may be modified or adaptively optimized (§§ 17, 18); preference and relational data (§ 11) or mixed and missing data may be handled (§ 7, 16); kernels, centroids, regression surfaces or principal component hyperplanes may be used, alternatively, as the class representatives (§§ 2, 4, 8, 13, 16); optimal scalings or discriminant spaces can be invoked (§§ 18, 9); multiple criteria problems and additional constraints (e.g. spatial, class size) are considered (§§ 2, 3, 20); the mixture problem and some continuous case theory is investigated in §§ 12, 16; contingency tables are treated in §§ 8, 19; hierarchies are relevant to § 22. Some applications concern real data problems from quality control, linguistics, picture processing, biology, and informatics (§§ 10, 14, 15, 23–25). Bibliographical notes and computer programs have been added, too.

However, no probabilistic results, no hints for choosing a special algorithm, and no techniques are given for graphical display, for interpreting the results, or assessing their relevance.

The book could be an excellent guide to the topics mentioned above and their development in France if its reading were not hindered by a huge number of misprints, a badly arranged lay-out and false or lacking reference possibilities (e.g. no index of topics). The mathematical rigour, the depth of argumentation and the notation is rather varying over the 25 chapters of the book due to the fact that 23 different authors contributed to this volume. Often the notation is obscure or only partially defined, ambiguous or heterogeneous (e.g. § 8 or p. 529 where, in the same line, for $k = i$ J_k and J_i mean different sets). Several chapters presuppose, without introducing the subject, results and notations from other French publications (e.g. on correspondence analysis). Substantial mathematical errors occur on page/line 185/16, 205/14, 339/14 and in the proofs on p. 373 and in § 18.3.7. A more concise and dense presentation would have been desirable.

However, the monograph will be useful for statisticians, data analysts and practitioners from many fields if they have a medium level of mathematical education, an elementary knowledge of cluster analysis and are willing to overcome the difficulties described above.

H.H. Bock

Address:

Prof. Dr. H.H. Bock
Institut f. Statistik u. Wirtschaftsmathematik
der RWTH Aachen
Wüllmerstr. 3, D-5100 Aachen

RICHMOND, Phyllis A.: *Introduction to PRECIS for North American Usage*. Littleton, CO: Libraries Unlimited 1981. 321 p., ISBN 0-87287-240-8

Das Buch beschränkt sich bei der Erklärung von PRECIS auf den englischsprachigen Bereich, insbesondere auf den Gebrauch in Nordamerika.

Es sei deshalb erwähnt, daß Versuche mit PRECIS auch in anderen Ländern mit anderen Sprachen durchgeführt werden. So wird PRECIS im französischen Sprachraum von verschiedenen Bibliotheken in Frankreich und der Bibliothèque Nationale du Québec erprobt.

Bei der Deutschen Bibliothek läuft derzeit ein einjähriges Projekt, in dem die Anwendung von PRECIS für die deutsche Sprache erprobt wird.

Dies sei vorangestellt, um zu verdeutlichen, daß das besprochene Buch nur einen Ausschnitt aus der PRECIS-Forschung darstellt.

Die Autorin schlägt vor, das Buch „Introduction to PRECIS for North American Usage“ gemeinsam mit dem PRECIS-Handbuch von Derek Austin zu benutzen. Die von ihr angeführten Beispiele und Erläuterungen gehen oft über das Handbuch hinaus. Die zahlreichen und übersichtlichen Verweisungen der Autorin ermöglichen es dem Leser, mit beiden Werken gleichzeitig zu arbeiten und aus den wertvollen Ergänzungen der Autorin zum Handbuch großen Nutzen zu ziehen.

Der Schwerpunkt liegt hierbei auf den syntaktischen Fragen, doch wird auch soweit auf die Semantik eingegangen, wie es für das Verständnis des Verweisungssystems in PRECIS erforderlich ist.