Diss. ETH Nr. 7418

PROBLEME UND ALGORITHMEN DER CLUSTERANALYSE

unter besonderer Berücksichtigung der Anwendung auf die landwirtschaftliche Typisierung

ABHANDLUNG

zur Erlangung

des Titels eines Doktors der technischen Wissenschaften

der

EIDGENOESSISCHEN TECHNISCHEN HOCHSCHULE ZUERICH

vorgelegt von

LENNART PIRKTL

Dipl. Math.

geboren am 27. September 1952

von Baden

Genehmigt auf Antrag von

Prof. Dr. D. Onigkeit, Referent

Prof. Dr. K. Hässig, Korreferent

Zürich, 1983

SUMMARY

The algorithms of Cluster Analysis have recently achieved their own distinct position within the group of multivariate techniques. This study aims to facilitate an appropriate choice of methods and to discuss capabilities and limitations of the procedures at hand.

The first of the eight chapters is an introduction, it shows the wide range of applications of automatic classification methods. Chapter two describes traditional classification in Agriculture. In addition previous work towards the establishment of the multivariate approach is discussed. Most classification procedures are based on the empirical measurement of dissimilarity between objects and between groups of objects. These and related conceptual and practical problems are treated in the third chapter. In chapter four hierarchical, iterative, density-seeking and heuristic algorithms are introduced. Some additional procedures are also considered. These are applied to the Ruspini-Data. Chapter five discusses the important concept of data transformations in the broad sense. In particular the conceptual deficiency of the widely spread z-transformation is demonstrated. The h-transformation $x \rightarrow x \div \mu$ is proposed as one of the means of removing some of the problems for the case of the ratio scale.

The counterbalancing function of the z-transformation is interpreted as the "Principle of equal contribution to distance measurement" ($E(Z_1-Z_2)^2$=const.). This principle is applied to the analysis of mixed data. The logarithmic transformation is then discussed, especially after calculation of ratios. Concerning the mean and variance of such ratios, the use of weighted moments is recommended. Chapter VI deals with the problem of interpreting clustering results, linked with the graphical representation of multivariate data: profiles, projections, dendrograms, structograms, stars, Chernoff-faces and Fourier-curves. Chapter VII compares the most important algorithms with reference to efficiency and computing requirements. For data sets with less than a thousand items relocation with optimization of the trace(W)-criterion is shown to be very effective. On the other hand, it had to be admitted that its dependance on the initial partition is a conceptual deficiency. In the case of larger data sets (up to a hundred thousand items) I propose a combination of quick and well performing algorithms. The concluding chapter includes three case studies, in which the implementations of the techniques of the previous sections are shown in practice.

In Appendix H some of the computer programs are listed, that were used on an IBM 3033 at the Computer Center of the Univerity of Zurich.

ZUSAMMENFASSUNG

Erst seit wenigen Jahren gewinnen Verfahren der Clusteranalyse als eigenständige Gruppe multivariater Verfahren stärker an Bedeutung. Es ist das Ziel dieser Arbeit, die Orientierung für eine sinnvolle Algorithmenwahl zu erleichtern sowie eine kritische Auseinandersetzung mit den Möglichkeiten und Grenzen dieser Verfahren zu führen.

Dazu gliedert sich die vorliegende Arbeit in acht Teile. Das erste Kapitel besitzt einführenden Charakter und gibt ferner einen Ueberblick der Wissenschaftsgebiete, in denen sich der clusteranalytische Ansatz durchsetzen konnte. Das zweite Kapitel beschäftigt sich mit konventionellen Klassifikationsverfahren im Landwirtschaftsbereich. Zudem würdigt es die bisher unternommenen Schritte zur Einführung multivariater Ansätze. Die überwiegende Mehrheit der Clustering-Algorithmen basiert auf der empirischen Bestimmung der Distanz zwischen Objekten resp. Objektgruppen. Die damit in Zusammenhang stehenden konzeptuellen und praktischen Probleme werden im dritten Kapitel behandelt. Im vierten Kapitel werden hierarchische, iterative, probabilistische, heuristische und einige weitere Algorithmen vorgestellt und am Datensatz von Ruspini ausgetestet. Einen wichtigen Platz nimmt Kapitel V ein. Es beinhaltet Datentransformationen im weiteren Sinne. Insbesondere zeigen wir konzeptuelle Mängel der weit verbreiteten z-Transformation sowie Ansätze zu deren Behebung auf. Hierzu ist bei ratioskalierten Daten die h-Transformation $x \rightarrow x^{\mu}$ zu zählen, welche die relative Streuung der Rohdaten beibehält.

Die gleichgewichtende Funktion der z-Transformation interpretieren wir als "Prinzip gleicher durchschnittlicher Distanzbeiträge" ($E(Z_1-Z_2)^2$=const.) und übertragen dieses zur Analyse gemischt-skalierter Merkmale auf binäre, ordinale und kategoriale Daten. Danach folgt eine Diskussion der Log-Transformation, insbesondere nach Quotientenbildung, für welche wir eine gewichtete Momentberechnung befürworten. In Kapitel VI bemühen wir uns um Interpretationsansätze von Clustering-Resultaten. Damit ist implizit die graphische Darstellung vieldimensionaler Daten angesprochen, wozu wir Verfahren wie Profile, Projektionen, Dendrogramme, Struktogramme, Sternfiguren, Chernoff-faces und Fourier-Reihen diskutieren und austesten. Das siebte Kapitel beschäftigt sich mit dem bewertenden Vergleich der wichtigsten Verfahren bezüglich der Effizienz bei der Klassenbildung und bezüglich des Rechenaufwandes beim Klassifikationsprozess. Bei unter tausend Merkmalsträgern erwies sich das auf dem trace(W)-Kriterium beruhende iterative Verfahren äusserst leistungsfähig, doch muss die Abhängigkeit von der Initialpartition als konzeptueller Mangel verstanden werden. Zur Klassifikation grösserer Datenmengen (tausend bis hunderttausend Objekte) wird die Kombination schneller und qualitativ hochstehender Verfahren vorgeschlagen. Im Schlusskapitel werden die gewonnenen Erkenntnisse an drei Fallstudien dargestellt.

In Anhang H sind einige der Computerprogramme aufgelistet, die auf der IBM 3033 des Rechenzentrums der Universität Zürich verwendet wurden.