Fuzzy Cluster Analysis

FUZZY CLUSTER ANALYSIS

METHODS FOR CLASSIFICATION DATA ANALYSIS AND IMAGE RECOGNITION

Frank Höppner
German Aerospace Center, Braunschweig, Germany
Frank Klawonn
University of Ostfriesland, Emden, Germany
Rudolf Kruse
University of Magdeburg, Germany
Thomas Runkler
Siemens AG, Munich, Germany

JOHN WILEY & SONS, LTD
Chichester • New York • Weinheim • Brisbane • Singapore • Toronto

Originally published in the German language by Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, D-65189 Wiesbaden, Germany, under the title "Frank Höppner/Frank Klawonn/Rudolf Kruse: Fuzzy-Clusteranalysen. 1. Auflage (1st Edition)". Copyright 1997 by Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden.

Baffins Lane, Chichester,
West Sussex, PO19 1UD, England

National 01243 779777
International (+44) 1243 779777

e-mail (for orders and customer service enquiries): cs-books@wiley.co.uk

Visit our Home Page on http://www.wiley.co.uk or http://www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons, Inc., 605 Third Avenue,
New York, NY 10158-0012, USA

Weinheim • Brisbane • Singapore • Toronto

Library of Congress Cataloging-in-Publication Data

Fuzzy-Clusteranalysen. English
Fuzzy cluster analysis : methods for classification, data analysis, and image recognition / Frank Höppner ... [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 0-471-98864-2 (cloth : alk. paper)
1. Cluster analysis. 2. Fuzzy sets. I. Höppner, Frank.
II. Title.
QA278.F8913 1999 99-25473
519.5'3—dc21 CIP

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 471 98864 2

Produced from camera-ready copy supplied by the authors
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

FUZZY CLUSTER ANALYSIS

Preface

When Lotfi Zadeh introduced the notion of a "fuzzy set" in 1965, his primary objective was to set up a formal framework for the representation and management of vague and uncertain knowledge. More than 20 years passed until fuzzy systems became established in industrial applications to a larger extent. Today, they are routinely applied especially in the field of control engineering. As a result of their success to translate knowledge-based approaches into a formal model that is also easy to implement, a great variety of methods for the usage of fuzzy techniques has been developed during the last years in the area of data analysis. Besides the possibility to take into account uncertainties within data, fuzzy data analysis allows us to learn a transparent and knowledge-based representation of the information inherent in the data. Areas of application for fuzzy cluster analysis include exploratory data analysis for pre-structuring data, classification and approximation problems, and the recognition of geometrical shapes in image processing.

When writing this book, our intention was to give a self-contained and methodical introduction to fuzzy cluster analysis with its areas of application and to provide a systematic description of different fuzzy clustering techniques, from which the user can choose the methods appropriate for his problem. The book applies to computer scientists, engineers and mathematicians in industry, research and teaching, who are occupied with data analysis, pattern recognition or image processing, or who take into consideration the application of fuzzy clustering methods in their area of work. Some basic knowledge in linear algebra is presupposed for the comprehension of the techniques and especially their derivation. Familiarity with fuzzy systems is not a requirement, because only in the chapter on rule generation with fuzzy clustering, more than the notion of a "fuzzy set" is necessary for understanding, and in addition, the basics of fuzzy systems are provided in that chapter.

Although this title is presented as a text book we have not included exercises for students, since it would not make sense to carry out the al-

gorithms by hand. We think that applying the algorithms to example data sets is the appropriate way to get a better understanding of the techniques. A software tool implementing most of the algorithms presented in chapters 1–5 and 7 together with the many example data sets discussed in this book are available as public domain software via the Internet at http://fuzzy.cs.uni-magdeburg.de/clusterbook/.

The book is an extension of a translation of our German book on fuzzy cluster analysis published by Vieweg Verlag in 1997. Most parts of the translation were carried out by Mark-Andre Krogel. The book would probably have appeared years later without his valuable support. The material of the book is partly based on lectures on fuzzy systems, fuzzy data analysis and fuzzy control that we gave at the Technical University of Braunschweig, at the University "Otto von Guericke" Magdeburg, at the University "Johannes Kepler" Linz, and at Ostfriesland University of Applied Sciences in Emden. The book is also based on a project in the framework of a research contract with Fraunhofer-Gesellschaft, on results from several industrial projects at Siemens Corporate Technology (Munich), and on joint work with Jim Bezdek at the University of West Florida. We thank Wilfried Euing and Hartmut Wolff for their advisory support during this project.

We would also like to express our thanks for the great support to Juliet Booker, Rhoswen Cowell, Peter Mitchell from Wiley and Reinald Klockenbusch from our German publisher Vieweg Verlag.

Frank Höppner
Frank Klawonn
Rudolf Kruse
Thomas Runkler

Contents

Introduction

For a fraction of a second, the receptors are fed with half a million items of data. Without any measurable time delay, those data items are evaluated and analysed, and their essential contents are recognized.

A glance at an image from TV or a newspaper, human beings are capable of this technically complex performance, which has not yet been achieved by any computer with comparable results. The bottleneck is no longer the optical sensors or data transmission, but the analysis and extraction of essential information. A single glance is sufficient for humans to identify circles and straight lines in accumulations of points and to produce an assignment between objects and points in the picture. Those points cannot always be assigned unambiguously to picture objects, although that hardly impairs human recognition performance. However, it is a big problem to model this decision with the help of an algorithm. The demand for an automatic analysis is high, though. Be it for the development of an autopilot for vehicle control, for visual quality control or for comparisons of large amounts of image data. The problem with the development of such a procedure is that humans cannot verbally reproduce their own procedures for image recognition, because it happens unconsciously. Conversely, humans have considerable difficulties recognizing relations in multi-dimensional data records that cannot be graphically represented. Here, they are dependent on computer supported techniques for data analysis, for which it is irrelevant whether the data consists of two- or twelve-dimensional vectors.

The introduction of fuzzy sets by L.A. Zadeh [104] in 1965 defined an object that allows the mathematical modelling of imprecise propositions. Since then this method has been employed in many areas to simulate how inferences are made by humans, or to manage uncertain information. This method can also be applied to data and image analysis.

Cluster analysis deals with the discovery of structures or groupings within data. Since hardly ever any disturbance or noise can be completely eliminated, some inherent data uncertainty cannot be avoided. That is

why fuzzy cluster analysis dispenses with unambiguous mapping of the data to classes and clusters, and instead computes degrees of membership that specify to what extend data belong to clusters.

The introductory chapter 1 relates fuzzy cluster analysis to the more general areas of cluster and data analysis, and provides the basic terminology. Here we focus on objective function models whose aim is to assign the data to clusters so that a given objective function is optimized. The objective function assigns a quality or error to each cluster arrangement, based on the distance between the data and the typical representatives of the clusters. We show how the objective function models can be optimized using an alternating optimization algorithm.

Chapter 2 is dedicated to fuzzy cluster analysis algorithms for the recognition of point-like clusters of different size and shape, which play a central role in data analysis.

The linear clustering techniques described in chapter 3 are suitable for the detection of clusters formed like straight lines, planes or hyperplanes, because of the suitable modification of the distance function that occurs in the objective functions. These techniques are appropriate for image processing, as well as for the construction of locally linear models of data with underlying functional interrelations.

Chapter 4 introduces shell clustering techniques, that aim to recognize geometrical contours such as borders of circles and ellipses by further modifications of the distance function. An extension of these techniques to non-smooth structures such as rectangles or other polygons is given in chapter 5.

The cluster estimation models described in chapter 6 abandon the objective function model. This allows handling of complex or not explicitly accessible systems, and leads to a generalized model with user-defined membership functions and prototypes.

Besides the assignment of data to classes, the determination of the number of clusters is a central problem in data analysis, which is also related to the more general problem of cluster validity. The aim of cluster validity is to evaluate whether clusters determined in an analysis are relevant or meaningful, or whether there might be no structure in the data that is covered by the clustering model. Chapter 7 provides an overview on cluster validity, and concentrates mainly on methods to determine the number of clusters, which are tailored to the different clustering algorithms.

Clusters can be interpreted as if-then rules. The structure information discovered by fuzzy clustering can therefore be translated to human readable fuzzy rule bases. The necessary techniques for this rule extraction are presented in chapter 8.

Readers who are interested in watching the algorithms at work can download free software via the Internet from http://fuzzy.cs.uni-magdeburg.de/clusterbook/.

Chapter 1

Basic Concepts

In everyday life, we often find statements like this:

> After a detailed analysis of the data available, we developed the opinion that the sales figures of our product could be increased by including the attribute *fuzzy* in the product's title.

Data analysis is obviously a notion which is readily used in everyday language. Everybody can understand it – however, there are different interpretations depending on the context. This is why these intuitive concepts like data, data analysis, cluster and partition have to be defined first.

1.1 Analysis of data

The concept of a datum is difficult to formalize. It originates from Latin and means "to be given". A datum is arbitrary information that makes an assertion about the state of a system, such as measurements, balances, degrees of popularity or On/Off states. We summarize the totality of all possible states, in which a system can be, under the concept *state* or *data space*. Any element of a data space describes a particular state of a system.

The data that has to be analysed may come from the area of medical diagnosis in the form of a database about patients, they may describe states of an industrial production plant, they may be available as time series which specify the progression of share prices, they may be obtained from statistical investigations, they may reflect opinions of experts, or they may be available as images.

Data analysis is always conducted to answer a particular question. That question implicitly determines the form of the answer: although it is dependent on the respective state of the system, it will always be of a particular *type*. Similarly, we want to summarize the possible answers to a question in a set that we call *result space*. In order to really gain information from the analysis, we require the result space to allow at least two different results. Otherwise, the answer would already unambiguously be given without any analysis.

In [5], data analysis is divided into four levels of increasing complexity. The first level consists of a simple frequency analysis, a reliability or credibility evaluation after which data identified as outliers are marked or eliminated, if necessary. On the second level, pattern recognition takes place, by which the data is grouped, and the groups are further structured, etc. These two levels are assigned to the area of exploratory data analysis, which deals with the investigation of data without assuming a mathematical model chosen beforehand that would have to explain the occurrence of the data and their structures. Figure 1.1 shows a set of data where an exploratory data analysis should recognize the two groups or clusters and assign the data to the respective groups.

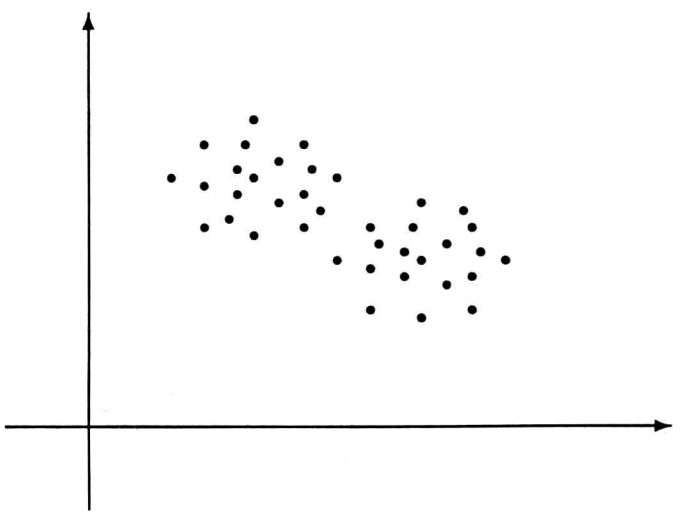

Figure 1.1: Recognition of two clusters by exploratory data analysis

On the third level of data analysis, the data are examined with respect to one or more mathematical models – for example in figure 1.1, whether the assumption is reasonable that the data are realizations of two two-

dimensional normally distributed random variables, and if so, what are the underlying parameters of the normal distributions. On the third level, a quantitative data analysis is usually performed, that means (functional) relations between the data should be recognized and specified if necessary, for instance by an approximation of the data using regression. In contrast, a purely qualitative investigation takes place on the second level, with the aim to group the data on the basis of a similarity concept.

Drawing conclusions and evaluating them is carried out on the fourth level. Here, conclusions can be predictions of future or missing data or an assignment to certain structures, for example, which pixels belong to the legs of a chair. An evaluation of the conclusions contains a judgement about how reliably the assignments can be made, whether modelling assumptions are realistic at all, etc. If necessary, a model that was constructed on the third level has to be revised.

The methods of fuzzy cluster analysis introduced in chapter 2 can essentially be categorized in the second level of data analysis, while the generation of fuzzy rules in chapter 8 belongs to the third level, because the rules serve as a description of functional relations. Higher order clustering techniques can also be assigned to the third level. Shell clustering, for example, not only aims at mapping of the data to geometrical contours such as circles, but is also used for a determination of parameters of geometrical contours, such as the circle's centre and radius.

Fuzzy clustering is a part of fuzzy data analysis that comprises two very different areas: the analysis of fuzzy data and the analysis of usual (crisp) data with the help of fuzzy techniques. We restrict ourselves mainly to the analysis of crisp data in the form of real-valued vectors with the help of fuzzy clustering methods. The advantages offered by a fuzzy assignment of data to groups in comparison to a crisp one will be clarified later on.

Even though measurements are usually afected by uncertainty, in most cases they provide concrete values so that fuzzy data are rarely obtained directly. An exception are public opinion polls that permit evaluations such as "very good" or "fairly bad" or, for instance, statements about time aspects such as "for quite a long time" or "for a rather short period of time". Statements like these correspond more to fuzzy sets than crisp values or intervals and should therefore be modelled with fuzzy sets. Methods to analyse fuzzy data like these are described in [6, 69, 73], among others. Another area, where fuzzy data are produced, is image processing. Grey values in grey scale pictures can be interpreted as degrees of membership to the colour black so that a grey scale picture represents a fuzzy set over the pixels. Even though we apply the fuzzy clustering techniques that are introduced in this book for image processing of black-and-white pictures only, these techniques can be extended to grey scale pictures by assigning

each pixel its grey value (transformed into the unit interval) as a weight. In this sense, fuzzy clustering techniques especially for image processing can be considered as methods to analyse fuzzy data.

1.2 Cluster analysis

Since the focus lies on fuzzy cluster analysis methods in this book, we can give only a short survey on general issues of cluster analysis. A more thorough treatment of this topic can be found in monographs such as [3, 16, 96].

The aim of a cluster analysis is to partition a given set of data or objects into clusters (subsets, groups, classes). This partition should have the following properties:

- Homogeneity within the clusters, i.e. data that belong to the same cluster should be as similar as possible.
- Heterogeneity between clusters, i.e. data that belong to different clusters should be as different as possible.

The concept of "similarity" has to be specified according to the data. Since the data are in most cases real-valued vectors, the Euclidean distance between data can be used as a measure of the dissimilarity. One should consider that the individual variables (components of the vector) can be of different relevance. In particular, the range of values should be suitably scaled in order to obtain reasonable distance values. Figures 1.2 and 1.3 illustrate this issue with a very simple example. Figure 1.2 shows four data points that can obviously be divided into the two clusters $\{x_1, x_2\}$ and $\{x_3, x_4\}$. In figure 1.3, the same data points are presented using a different scale where the units on the x-axis are closer together while they are more distant on the y-axis. The effect would be even stronger if one would take kilo-units for the x-axis and milli-units for the y-axis. Two clusters can be recognized in figure 1.3, too. However, they combine the data point x_1 with x_4 and x_2 with x_3, respectively.

Further difficulties arise when not only real-valued variables occur but also integer-valued ones or even abstract classes (e.g. types of cars: convertible, sedan, truck etc.). Of course, the Euclidean distance can be computed for integer values. However, the integer values in a variable can produce a cluster partition where a cluster is simply assigned to each occurring integer number. That can be meaningful or completely undesirable dependent on the data and the question to be investigated. Numbers can be assigned to abstract classes, and thus the Euclidean distance can be applied again.