Advances in Fuzzy Clustering and its Applications

EDITED BY

J. Valente de Oliveira
University of Algarve, Portugal

W. Pedrycz
University of Alberta, Canada
Systems Research Institute of the Polish Academy of Sciences, Poland

John Wiley & Sons, Ltd
Advances in Fuzzy Clustering and its Applications
Advances in Fuzzy Clustering and its Applications

EDITED BY

J. Valente de Oliveira
University of Algarve, Portugal

W. Pedrycz
University of Alberta, Canada
Systems Research Institute of the Polish Academy of Sciences, Poland
Contents

List of Contributors xi

Foreword xv

Preface xvii

Part I Fundamentals 1

1 Fundamentals of Fuzzy Clustering 3
   Rudolf Kruse, Christian Döring and Marie-Jeanne Lesot
   1.1 Introduction 3
   1.2 Basic Clustering Algorithms 4
   1.3 Distance Function Variants 14
   1.4 Objective Function Variants 18
   1.5 Update Equation Variants: Alternating Cluster Estimation 25
   1.6 Concluding Remarks 27
   Acknowledgements 28
   References 29

2 Relational Fuzzy Clustering 31
   Thomas A. Runkler
   2.1 Introduction 31
   2.2 Object and Relational Data 31
   2.3 Object Data Clustering Models 34
   2.4 Relational Clustering 38
   2.5 Relational Clustering with Non-spherical Prototypes 41
   2.6 Relational Data Interpreted as Object Data 45
   2.7 Summary 46
   2.8 Experiments 46
   2.9 Conclusions 49
   References 50

3 Fuzzy Clustering with Minkowski Distance Functions 53
   Patrick J.F. Groenen, Uzay Kaymak and Joost van Rosmalen
   3.1 Introduction 53
   3.2 Formalization 54
   3.3 The Majorizing Algorithm for Fuzzy C-means with Minkowski Distances 56
8 Fuzzy Clustering of Fuzzy Data
Pierpaolo D’Urso

8.1 Introduction 155
8.2 Informational Paradigm, Fuzziness and Complexity in Clustering Processes 156
8.3 Fuzzy Data 160
8.4 Fuzzy Clustering of Fuzzy Data 165
8.5 An Extension: Fuzzy Clustering Models for Fuzzy Data Time Arrays 176
8.6 Applicative Examples 180
8.7 Concluding Remarks and Future Perspectives 187
References 189

9 Inclusion-based Fuzzy Clustering
Samia Nefti-Meziani and Mourad Oussalah

9.1 Introduction 193
9.2 Background: Fuzzy Clustering 195
9.3 Construction of an Inclusion Index 196
9.4 Inclusion-based Fuzzy Clustering 198
9.5 Numerical Examples and Illustrations 201
9.6 Conclusions 206
Acknowledgements 206
Appendix 9A.1 207
References 208

10 Mining Diagnostic Rules Using Fuzzy Clustering
Giovanna Castellano, Anna M. Fanelli and Corrado Mencar

10.1 Introduction 211
10.2 Fuzzy Medical Diagnosis 212
10.3 Interpretability in Fuzzy Medical Diagnosis 213
10.4 A Framework for Mining Interpretable Diagnostic Rules 216
10.5 An Illustrative Example 221
10.6 Concluding Remarks 226
References 226

11 Fuzzy Regression Clustering
Mikal Sato-Ilic

11.1 Introduction 229
11.2 Statistical Weighted Regression Models 230
11.3 Fuzzy Regression Clustering Models 232
11.4 Analyses of Residuals on Fuzzy Regression Clustering Models 237
CONTENTS  ix

15.4 Applications  324
15.5 Future Perspectives and Conclusions  331
Acknowledgement  331
References  331

16 Fuzzy Clustering of Parallel Data Streams  333
Jürgen Beringer and Eyke Hüllermeier

16.1 Introduction  333
16.2 Background  334
16.3 Preprocessing and Maintaining Data Streams  336
16.4 Fuzzy Clustering of Data Streams  340
16.5 Quality Measures  343
16.6 Experimental Validation  345
16.7 Conclusions  350
References  351

17 Algorithms for Real-time Clustering and Generation of Rules from Data  353
Dimitar Filev and Plamen Angelov

17.1 Introduction  353
17.2 Density-based Real-time Clustering  355
17.3 FSPC: Real-time Learning of Simplified Mamdani Models  358
17.4 Applications  362
17.5 Conclusion  367
References  368

Part V Applications and Case Studies  371

18 Robust Exploratory Analysis of Magnetic Resonance Images using FCM with Feature Partitions  373
Mark D. Alexiuk and Nick J. Pizzi

18.1 Introduction  373
18.2 FCM with Feature Partitions  374
18.3 Magnetic Resonance Imaging  379
18.4 FMRI Analysis with FCMP  381
18.5 Data-sets  382
18.6 Results and Discussion  384
18.7 Conclusion  390
Acknowledgements  390
References  390

19 Concept Induction via Fuzzy C-means Clustering in a High-dimensional Semantic Space  393
Dawei Song, Guihong Cao, Peter Bruza and Raymond Lau

19.1 Introduction  393
19.2 Constructing a High-dimensional Semantic Space via Hyperspace Analogue to Language  395
19.3 Fuzzy C-means Clustering  397
List of Contributors

EDITORS

José Valente de Oliveira  
The UAlg Informatics Lab  
Faculty of Science and Technology  
University of Algarve  
Portugal

Witold Pedrycz  
Department of Electrical and Computer Engineering  
University of Alberta, Canada  
and Systems Research Institute of the Polish Academy of Sciences Warsaw, Poland

AUTHORS

János Abonyi  
University of Veszprém  
Department of Process Engineering  
Hungary

Benjamin Bird  
School of Chemistry  
University of Nottingham  
UK

Mark D. Alexiuk  
Department of Electrical and Computer Engineering  
University of Manitoba  
Canada

Peter Bruza  
School of Information Technology  
Queensland University of Technology  
Australia

Palmen Angelov  
Department of Communication Systems  
Lancaster University  
UK

Guihong Cao  
Department d’Informatique et Recherche operationnelle  
Université de Montreal  
Canada

Jürgen Beringer  
Fakultät für Informatik  
Otto-von-Guericke-Universität Magdeburg  
Germany

Giovanna Castellano  
Department of Computer Science  
University of Bari  
Italy

Michael R. Berthold  
Department of Computer and Information Science  
University of Konstanz  
Germany

Mario G.C.A. Cimino  
Dipartimento di Ingegneria dell’Informazione: Elettronica, Informatica, Telecomunicazioni  
University of Pisa  
Italy
LIST OF CONTRIBUTORS

Christian Döring
School of Computer Science
Otto-von-Guericke-Universität Magdeburg
Germany

Fernando Gomide
State University of Campinas
Brazil

Pierpaolo D’Urso
Dipartimento di Scienze Economiche
Gestionali e Sociali
Università degli Studi del Molise
Italy

Patrick J.F. Groenen
Econometric Institute
Erasmus University Rotterdam
The Netherlands

Anna M. Fanelli
Department of Computer Science
University of Bari
Italy

Eyke Hüllermeier
Fakultät für Informatik
Otto-von-Guericke-Universität Magdeburg
Germany

Balázs Feil
University of Veszprém
Department of Process Engineering
Hungary

Uzay Kaymak
Econometric Institute
Erasmus University Rotterdam
The Netherlands

Dimitar Filev
Ford Motor Company
Dearborn
USA

Rudolf Kruse
School of Computer Science
Otto-von-Guericke-Universität Magdeburg
Germany

Hichem Frigui
Department of Computer Engineering
and Computer Science
University of Louisville
USA

Raymond Lau
Department of Information Systems
City University of Hong Kong
Hong Kong SAR

Jonathan M. Garibaldi
School of Computer Science and Information
Technology
University of Nottingham
UK

Beatrice Lazzerini
Dipartimento di Ingegneria dell’Informazione: Elettronica, Informatica, Telecomunicazioni
University of Pisa
Italy

Mike W. George
School of Chemistry
University of Nottingham
UK

Marie-Jeanne Lesot
School of Computer Science
Otto-von-Guericke-Universität Magdeburg
Germany

Joydeep Ghosh
Department of Electrical and
Computer Engineering
University of Texas at Austin
Texas
USA

Francesco Marcelloni
Dipartimento di Ingegneria dell’Informazione: Elettronica, Informatica, Telecomunicazioni
University of Pisa
Italy
Corrado Mencar  
Department of Computer Science  
University of Bari  
Italy

Samia Nefti-Meziani  
Department of Engineering and Technology  
University of Manchester  
UK

Mourad Oussalah  
Electronics, Electrical and Computing Engineering  
The University of Birmingham  
UK

David E. Patterson  
Department of Computer and Information Science  
University of Konstanz  
Germany

Nick J. Pizzi  
National Research Council Institute for Biodiagnostics  
Winnipeg  
Canada

Kunal Punera  
Department of Electrical and Computer Engineering  
University of Texas at Austin  
USA

Thomas A. Runkler  
Siemens AG Corporate Technology Information and Communications  
München  
Germany

Mika Sato-Ilic  
Faculty of Systems and Information Engineering  
University of Tsukuba  
Japan

Leila Roling Scariot da Silva  
State University of Campinas  
Brazil

Dawei Song  
Knowledge Media Institute  
The Open University Milton Keynes  
UK

George E. Tsekouras  
Department of Cultural Technology and Communication University of the Aegean  
Mytilene  
Greece

Joost van Rosmalen  
Econometric Institute  
Erasmus University Rotterdam  
The Netherlands

Xiao-Ying Wang  
School of Computer Science and Information Technology  
University of Nottingham  
UK

Richard Weber  
Department of Industrial Engineering  
University of Chile Santiago  
Chile

Bernd Wiswedel  
Department of Computer and Information Science  
University of Konstanz  
Germany

Ronald Yager  
Iona College  
New Rochelle  
USA
Well, here I am writing a foreword for this book. Here is the (free dictionary, Farlex) definition:

‘foreword - a short introductory essay preceding the text of a book.’

An essay about fuzzy clustering? For inspiration, I looked at the forewords in my first two books. When I wrote my first book about fuzzy clustering (Bezdek, 1981), I asked Lotfi Zadeh to write a foreword for it. By then, Lotfi and I were friends, so he did it, and I was happy. But why? Was it to prove to you that I could get him to do it? Was it because he would say things that had never been said about fuzzy models? Was it a promotional gimmick that the publisher thought would get more buyers interested? Was it . . . hmmm, I still didn’t know, so I read more carefully.

Lotfi speculated on a variety of possibilities for fuzzy clustering in that foreword. The most interesting sentence (Bezdek, 1981, p. 5) was perhaps:

“Although the results of experimental studies reported in this book indicate that fuzzy clustering techniques often have significant advantages over more conventional methods, universal acceptance of the theory of fuzzy sets as a natural basis for pattern recognition and cluster analysis is not likely to materialize in the very near future.”

In short, his foreword was careful, and it was cautionary – Lotfi speculated that fuzzy clustering might not assume a central place in clustering, but this seems overshadowed by his more general worry about the role of fuzzy models in computation.

My second book (Bezdek and Pal, 1992) was much more similar to this volume than my first, because the 1981 effort was a one-author text, while the 1992 book was a collection of 51 papers (the “chapters”) that Pal and I put together (we were editors, just like de Oliveira and Pedrycz) that seemed to provide a state-of-the-art “survey” of what was happening with fuzzy models in various pattern recognition domains in 1992. Perhaps the principal difference between these two books is that fuzzy clustering was only one of the five topics of our 1992 book, whereas the current volume is only about fuzzy clustering. The other noticeable difference was that the papers we collected had already been published elsewhere, whereas the chapters in this book have not.

I am looking at the foreword to our 1992 book right now, again written by Lotfi. Well, a lot of positive things happened for fuzzy sets in the 11 years that separated these two forewords (read, Japan builds fuzzy controllers), and Lotfi’s 1992 foreword was both more historical and more confident than the 1981 offering. Here is the first sentence of that 1992 forward:

“To view the contents of this volume in a proper perspective it is of historical interest to note that the initial development of the theory of fuzzy sets was motivated in large measure by problems in pattern recognition and cluster analysis.”
Did you notice that Lotfi used exactly the same term “pattern recognition and cluster analysis” in both forewords? In contradistinction, I believe that most people today view clustering as one of many topics encompassed by the much broader field of pattern recognition (classifier design, feature selection, image processing, and so on). My guess is that Lotfi probably used the term pattern recognition almost as a synonym for classification. This is a small point, but in the context of this volume, an interesting one, because to this day, Lotfi contends that the word cluster is ill defined, and hence cluster analysis is not really a topic at all. Nonetheless, you have in your hands a new book about fuzzy cluster analysis.

What should I point out to you in 2006 about this topic? Well, the main point is that fuzzy clustering is now a pretty mature field. I just “googled” the index term “fuzzy cluster analysis,” and the search returned this statistic at 1 p.m. on September 6, 2006:

“Results 1–10 of about 1 640 000 for fuzzy cluster analysis (0.34 seconds).”

Never mind duplication, mixed indexing, and all the other false positives represented by this statistic. The fact is fuzzy clustering is a pretty big field now. There are still some diehard statisticians out there who deny its existence, much less its value to real applications, but by and large this is no longer a controversial undertaking, nor is its real value to practitioners questionable. Moreover, I can pick any chapter in this book and get returns from Google that amaze me. Example: Chapter 4 has the somewhat exotic title “Fuzzy Clustering with Minkowski Distance Functions.” What would you guess for this topic – 12 papers? Here is the return:

“Results 1–10 of about 20 000 for Fuzzy Clustering with Minkowski distance functions (0.37 seconds).”

There aren’t 20 000 papers out there about this topic, but there are probably a few hundred, and this is what makes the current book useful. Most of these chapters offer an encapsulated survey of (some of) the most important work on their advertised contents. This is valuable, because I don’t want to sift through 20 000 entries to find the good stuff about Minkowski-based fuzzy clustering – I want the experts to guide me to 20 or 30 papers that have it.

Summary. We no longer need worry whether the topics in this fuzzy clustering book are good stuff – they are. What we need that these chapters provide is a quick index to the good stuff. And for this, you should be grateful (and buy the book, for which de Oliveira and Pedrycz will be grateful!), because if you rely on “google,” you can spend the rest of your life sifting through the chaff to find the grain.

Jim BezdekJim Bezdek
Pensacola, USA
Preface

Clustering has become a widely accepted synonym of a broad array of activities of exploratory data analysis and model development in science, engineering, life sciences, business and economics, defense, and biological and medical disciplines. Areas such as data mining, image analysis, pattern recognition, modeling, and bio-informatics are just tangible examples of numerous pursuits that vigorously exploit the concepts and algorithms of clustering treated as essential tools for problem formulation and development of specific solutions or a vehicle facilitating interpretation mechanisms. The progress in the area happens at a high pace and these developments concern the fundamentals, algorithmic enhancements, computing schemes, and validation practices. The role of fuzzy clustering becomes quite prominent within the general framework of clustering. This is not surprising given the fact that clustering helps gain an interesting insight into data structure, facilitate efficient communication with users and data analysts, and form essential building blocks for further modeling pursuits. The conceptual underpinnings of fuzzy sets are particularly appealing, considering their abilities to quantify a level of membership of elements to detected clusters that are essential when dealing with the inherent phenomenon of partial belongingness to the group. This feature is of particular interest when dealing with various interpretation activities.

Even a very quick scan of the ongoing research reveals how dynamic the area of fuzzy clustering really is. For instance, a simple query on Science Direct “fuzzy clustering” returns slightly under 400 hits (those are the papers published since 2000). A similar search on ISI Web of Knowledge returns more than 500 hits. In IEEE Xplore one can find around 800 hits. More than half of these entries have been published after 2000. These figures offer us an impression about the rapid progress in the area and highlight a genuine wealth of the applications of the technology of fuzzy clustering.

This volume aims at providing a comprehensive, coherent, and in depth state-of-the-art account on fuzzy clustering. It offers an authoritative treatment of the subject matters presented by leading researchers in the area. While the volume is self-contained by covering some fundamentals and offering an exposure to some preliminary material on algorithms and practice of fuzzy clustering, it offers a balanced and broad coverage of the subject including theoretical fundamentals, methodological insights, algorithms, and case studies.

The content of the volume reflects the main objectives we intend to accomplish. The organization of the overall material helps the reader to proceed with some introductory material, move forward with more advanced topics, become familiar with recent algorithms, and finally gain a detailed knowledge of various application-driven facets.

The contributions have been organized into five general categories: Fundamentals, Visualization, Algorithms and Computational Aspects, Real-time and Dynamic Clustering, and Applications and Case Studies. They are fairly reflective of the key pursuits in the area.

Within the section dealing with the fundamentals, we are concerned with the principles of clustering as those are seen from the perspective of fuzzy sets. We elaborate on the role of fuzzy sets in data analysis, discuss the principles of data organization, and present fundamental algorithms and their augmentations. Different paradigms of unsupervised learning along with so-called knowledge-based clustering and data organization are also addressed in detail. This part is particularly aimed at the readers who would intend to gather some background material and have a quick yet carefully organized look at the essential of the methodology of fuzzy clustering.
In fuzzy clustering, visualization is an emerging subject. Due to its huge potential to address interpretation and validation issues visualization deserves to be treated as a separate topic.

The part entitled Algorithms and Computational Aspects focuses on the major lines of pursuits on the algorithmic and computational augmentations of fuzzy clustering. Here the major focus is on the demonstration of effectiveness of the paradigm of fuzzy clustering in high-dimensional problems, distributed problem solving, and uncertainty management.

The chapters arranged in the group entitled Real-time and Dynamic Clustering describe the state-of-the-art algorithms for dynamical developments of clusters, i.e., for clustering built for data gathered over time. Since new observations are available at each time instant, a dynamic update of clusters is required.

The Applications and Case Studies part is devoted to a series of applications in which fuzzy clustering plays a pivotal role. The primary intent is to discuss its role in the overall design process in various tasks of prediction, classification, control, and modeling. Here it becomes highly instructive to highlight at which phase of the design clustering is of relevance, what role it plays, and how the results – information granules – facilitate further detailed development of models or enhance interpretation aspects.

**PART I FUNDAMENTALS**

The part on Fundamentals consists of four chapters covering the essentials of fuzzy clustering and presenting a rationality and a motivation, basic algorithms and their various realizations, and cluster validity assessment.

Chapter 1 starts with an introduction to basic clustering algorithms including hard, probabilistic, and possibilistic ones. Then more advanced methods are presented, including the Gustafson–Kessel algorithm and kernel-based fuzzy clustering. Variants on a number of algorithm components as well as on problem formulations are also considered.

Chapter 2 surveys the most relevant methods of relational fuzzy clustering, i.e., fuzzy clustering for relational data. A distinction between object and relational data is presented and the consequences of this distinction on clustering algorithms are thoroughly analyzed. A most useful taxonomy for relational clustering algorithms together with some guidelines for selecting clustering schemes for a given application can also be found in this chapter.

In Chapter 3 the authors offer a contribution that deals with another fundamental issue in clustering: distance functions. The focus is on fuzzy clustering problems and algorithms using the Minkowski distance – definitely an interesting and useful idea.

In Chapter 4 the authors discuss the combination of multiple partitioning obtained from independent clustering runs into a consensus partition – a topic that is gaining interest and importance. A relevant review of commonly used approaches, new consensus strategies (including one based on information-theoretic $K$-means), as well as a thorough experimental evaluation of these strategies are presented.

**PART II VISUALIZATION**

Visualization is an important tool in data analysis and interpretation. Visualization offers the user the possibility of quickly inspecting a huge volume of data, and quickly selecting data space regions of interest for further analysis. Generally speaking, this is accomplished by producing a low-dimensional graphical representation of the clusters. The part of the book on Visualization consists of two major contributions.

Chapter 5 reviews relevant approaches to validity and visualization of clustering results. It also presents novel tools that allow the visualization of multi-dimensional data points in terms of bi-dimensional plots which facilitates the assessment of clusters’ goodness. The chapter ends with an appendix with a comprehensive description of cluster validity indexes.

Chapter 6 aims at helping the user to visually explore clusters. The approach consists of the construction of local, one-dimensional neighborhood models, the so-called neighborgrams. An algorithm is
included that generates a subset of neighborgrams from which the user can manage potential cluster candidates during the clustering process. This can be viewed as a form of integrating user domain knowledge into the clustering process.

**PART III  ALGORITHMS AND COMPUTATIONAL ASPECTS**

This part provides the major lines of work on algorithmic and computational augmentations of fuzzy clustering with the intention of demonstrating its effectiveness in high-dimensional problems, distributed problem solving and uncertainty handling. Different paradigms of unsupervised learning along with so-called knowledge-based clustering and data organization are also addressed.

Chapter describes and evaluates a clustering algorithm based on the Yager’s participatory learning rule. This learning rule pays special attention to current knowledge as it dominates the way in which new data are used for learning. In participatory clustering the number of clusters is not given a priori as it depends on the cluster structure that is dynamically built by the algorithm.

Chapter 8 offers a comprehensive and in-depth study on fuzzy clustering of fuzzy data. The authors of Chapter 9 also address the problem of clustering fuzzy data. In this case, clustering is based on the amount of mutual inclusion between fuzzy sets, especially between data and cluster prototypes.

Extraction of semantically valid rules from data is an active interdisciplinary research topic with foundations in computer and cognitive sciences, psychology, and philosophy. Chapter 10 addresses this topic from the clustering perspective. The chapter describes a clustering framework for extracting interpretable rules for medical diagnostics.

Chapter 11 focuses on the combination of regression models with fuzzy clustering. The chapter describes and evaluates several regression models for updating the partition matrix in clustering algorithms. The evaluation includes an analysis of residuals and reveals the interesting characteristics of this class of algorithm.

Hierarchical fuzzy clustering is discussed in Chapter 12. The chapter presents a clustering-based systematic approach to fuzzy modeling that takes into account the following three issues: (1) the number of clusters required a priori in fuzzy clustering; (2) initialization of fuzzy clustering methods, and (3) the trade off between accuracy and interpretability.

Chapter 13 deals with the process of inferring dissimilarity relations from data. For this, two methods are analyzed with respect to factors such as generalization and computational complexity. The approach is particularly interesting for applications where the nature of dissimilarity is conceptual rather than metric.

Chapter 14 describes how clustering and feature selection can be unified to improve the discovery of more relevant data structures. An extension of the proposed algorithm for dealing with an unknown number of clusters is also presented. Interesting applications on image segmentation and text categorization are included.

**PART IV  REAL-TIME AND DYNAMIC CLUSTERING**

Real-time and dynamic clustering deals with clustering with time-varying or noisy data and finds its applications in areas as distinct as video or stock market analysis. Three chapters focus on this timely topic.

Chapter 15 provides a review of dynamic clustering emphasizing its relationship with the area of data mining. Data mining is a matter of paramount relevance today and this chapter shows how dynamic clustering can be brought into the picture. The chapter also describes two novel approaches to dynamic clustering.

Chapter 16 describes the development of an efficient online version of the fuzzy C-means clustering for data streams, i.e., data of potentially unbound size whose continuous evolution is not under the control of the analyzer.
Chapter 17 presents two approaches to real-time clustering and generation of rules from data. The first approach concerns a density-driven approach with its origin stemming from the techniques of mountain and subtractive clustering while the second one looks at the distance based with foundations in the $k$-nearest neighbors and self-organizing maps.

**PART V APPLICATIONS AND CASE STUDIES**

The last part of the book includes three chapters describing various applications and interesting case studies in which fuzzy clustering plays an instrumental role. The function of fuzzy clustering is discussed in the overall design process in a variety of tasks such as prediction, classification, and modeling.

Chapter 18 presents a novel clustering algorithm that incorporates spatial information by defining multiple feature partitions and shows its application to the analysis of magnetic resonance images.

Chapter 19 exploits both the $K$-means and the fuzzy C-means clustering algorithms as the means to identify correlations between words in texts, using the hyperspace analogue to language (HAL) model.

Another bio-medical application is provided in Chapter 20 where fuzzy clustering techniques are used in the identification of cancerous cells.

**FINAL REMARKS**

All in all, fuzzy clustering forms a highly enabling technology of data analysis. The area is relatively mature and exhibits a rapid expansion in many different directions including a variety of new concepts, methodologies, algorithms, and innovative and highly advanced applications.

We do hope that the contributions compiled in this volume will bring the reader a fully updated and highly comprehensive view of the recent developments in the fundamentals, algorithms, and applications of fuzzy clustering.

Our gratitude goes to all authors for sharing their expertise and recent research outcomes and reviewers whose constructive criticism was of immense help in producing a high quality volume. Finally, our sincere thanks go to the dedicated and knowledgeable staff at John Wiley & Sons, Ltd, who were highly instrumental in all phases of the project.
Part I
Fundamentals
1 Fundamentals of Fuzzy Clustering

Rudolf Kruse, Christian Döring, and Marie-Jeanne Lesot

Department of Knowledge Processing and Language Engineering, University of Magdeburg, Germany

1.1 INTRODUCTION

Clustering is an unsupervised learning task that aims at decomposing a given set of objects into subgroups or clusters based on similarity. The goal is to divide the data-set in such a way that objects (or example cases) belonging to the same cluster are as similar as possible, whereas objects belonging to different clusters are as dissimilar as possible. The motivation for finding and building classes in this way can be manifold (Bock, 1974). Cluster analysis is primarily a tool for discovering previously hidden structure in a set of unordered objects. In this case one assumes that a ‘true’ or natural grouping exists in the data. However, the assignment of objects to the classes and the description of these classes are unknown. By arranging similar objects into clusters one tries to reconstruct the unknown structure in the hope that every cluster found represents an actual type or category of objects. Clustering methods can also be used for data reduction purposes. Then it is merely aiming at a simplified representation of the set of objects which allows for dealing with a manageable number of homogeneous groups instead of with a vast number of single objects. Only some mathematical criteria can decide on the composition of clusters when classifying data-sets automatically. Therefore clustering methods are endowed with distance functions that measure the dissimilarity of presented example cases, which is equivalent to measuring their similarity. As a result one yields a partition of the data-set into clusters regarding the chosen dissimilarity relation.

All clustering methods that we consider in this chapter are partitioning algorithms. Given a positive integer \( c \), they aim at finding the best partition of the data into \( c \) groups based on the given dissimilarity measure and they regard the space of possible partitions into \( c \) subsets only. Therein partitioning clustering methods are different from hierarchical techniques. The latter organize data in a nested sequence of groups, which can be visualized in the form of a dendrogram or tree. Based on a dendrogram one can decide on the number of clusters at which the data are best represented for a given purpose. Usually the number of (true) clusters in the given data is unknown in advance. However, using the partitioning methods one is usually required to specify the number of clusters \( c \) as an input parameter. Estimating the actual number of clusters is thus an important issue that we do not leave untouched in this chapter.
A common concept of all described clustering approaches is that they are prototype-based, i.e., the clusters are represented by cluster prototypes $C_i$, $i = 1, \ldots, c$. Prototypes are used to capture the structure (distribution) of the data in each cluster. With this representation of the clusters we formally denote the set of prototypes $C = \{C_1, \ldots, C_c\}$. Each prototype $C_i$ is an $n$-tuple of parameters that consists of a cluster center $c_i$ (location parameter) and maybe some additional parameters about the size and the shape of the cluster. The cluster center $c_i$ is an instantiation of the attributes used to describe the domain, just as the data points in the data-set to divide. The size and shape parameters of a prototype determine the extension of the cluster in different directions of the underlying domain. The prototypes are constructed by the clustering algorithms and serve as prototypical representations of the data points in each cluster.

The chapter is organized as follows: Section 1.2 introduces the basic approaches to hard, fuzzy, and possibilistic clustering. The objective function they minimize is presented as well as the minimization method, the alternating optimization (AO) scheme. The respective partition types are discussed and special emphasis is put on a thorough comparison between them. Further, an intuitive understanding of the general properties that distinguish their results is presented. Then a systematic overview of more sophisticated fuzzy clustering methods is presented. In Section 1.3, the variants that modify the used distance functions for detecting specific cluster shapes or geometrical contours are discussed. In Section 1.4 variants that modify the optimized objective functions for improving the results regarding specific requirements, e.g., dealing with noise, are reviewed. Lastly, in Section 1.5, the alternating cluster estimation framework is considered. It is a generalization of the AO scheme for cluster model optimization, which offers more modeling flexibility without deriving parameter update equations from optimization constraints. Section 1.6 concludes the chapter pointing at related issues and selected developments in the field.

1.2 BASIC CLUSTERING ALGORITHMS

In this section, we present the fuzzy $C$-means and possibilistic $C$-means, deriving them from the hard $c$-means clustering algorithm. The latter one is better known as $k$-means, but here we call it (hard) $C$-means to unify the notation and to emphasize that it served as a starting point for the fuzzy extensions. We further restrict ourselves to the simplest form of cluster prototypes at first. That is, each prototype only consists of the center vectors, $C_i = (c_i)$, such that the data points assigned to a cluster are represented by a prototypical point in the data space. We consider as a distance measure $d$ an inner product norm induced distance as for instance the Euclidean distance. The description of the more complex prototypes and other dissimilarity measures is postponed to Section 1.3, since they are extensions of the basic algorithms discussed here.

All algorithms described in this section are based on objective functions $J$, which are mathematical criteria that quantify the goodness of cluster models that comprise prototypes and data partition. Objective functions serve as cost functions that have to be minimized to obtain optimal cluster solutions. Thus, for each of the following cluster models the respective objective function expresses desired properties of what should be regarded as “best” results of the cluster algorithm. Having defined such a criterion of optimality, the clustering task can be formulated as a function optimization problem. That is, the algorithms determine the best decomposition of a data-set into a predefined number of clusters by minimizing their objective function. The steps of the algorithms follow from the optimization scheme that they apply to approach the optimum of $J$. Thus, in our presentation of the hard, fuzzy, and possibilistic $c$-means we discuss their respective objective functions first. Then we shed light on their specific minimization scheme.

The idea of defining an objective function and have its minimization drive the clustering process is quite universal. Aside from the basic algorithms many extensions and modifications have been proposed that aim at improvements of the clustering results with respect to particular problems (e.g., noise, outliers). Consequently, other objective functions have been tailored for these specific applications. We address the most important of the proposed objective function variants in Section 1.4. However, regardless of the specific objective function that an algorithm is based on, the objective function is a goodness measure.
Thus it can be used to compare several clustering models of a data-set that have been obtained by the same algorithm (holding the number of clusters, i.e., the value of \( c \), fixed).

In their basic forms the hard, fuzzy, and possibilistic C-means algorithms look for a predefined number of \( c \) clusters in a given data-set, where each of the clusters is represented by its center vector. However, hard, fuzzy, and possibilistic C-means differ in the way they assign data to clusters, i.e., what type of data partitions they form. In classical (hard) cluster analysis each datum is assigned to exactly one cluster. Consequently, the hard C-means yield exhaustive partitions of the example set into non-empty and pairwise disjoint subsets. Such hard (crisp) assignment of data to clusters can be inadequate in the presence of data points that are almost equally distant from two or more clusters. Such special data points can represent hybrid-type or mixture objects, which are (more or less) equally similar to two or more types. A crisp partition arbitrarily forces the full assignment of such data points to one of the clusters, although they should (almost) equally belong to all of them. For this purpose the fuzzy clustering approaches presented in Sections 1.2.2 and 1.2.3 relax the requirement that data points have to be assigned to one (and only one) cluster. Data points can belong to more than one cluster and even with different degrees of membership to the different clusters. These gradual cluster assignments can reflect present cluster structure in a more natural way, especially when clusters overlap. Then the memberships of data points at the overlapping boundaries can express the ambiguity of the cluster assignment.

The shift from hard to gradual assignment of data to clusters for the purpose of more expressive data partitions founded the field of fuzzy cluster analysis. We start our presentation with the hard C-means and later on we point out the relatedness to the fuzzy approaches that is evident in many respects.

### 1.2.1 Hard c-means

In the classical C-means model each data point \( x_j \) in the given data-set \( X = \{x_1, \ldots, x_n\}, \ X \subseteq \mathbb{R}^p \) is assigned to exactly one cluster. Each cluster \( \Gamma_i \) is thus a subset of the given data-set, \( \Gamma_i \subset X \). The set of clusters \( \Gamma = \{\Gamma_1, \ldots, \Gamma_c\} \) is required to be an exhaustive partition of the data-set \( X \) into \( c \) non-empty and pairwise disjoint subsets \( \Gamma_i, 1 < c < n \). In the C-means such a data partition is said to be optimal when the sum of the squared distances between the cluster centers and the data points assigned to them is minimal (Krishnapuram and Keller, 1996). This definition follows directly from the requirement that clusters should be as homogeneous as possible. Hence the objective function of the hard C-means can be written as follows:

\[
J_h(X, U_h, C) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} d_{ij},
\]  

(1.1)

where \( C = \{C_1, \ldots, C_c\} \) is the set of cluster prototypes, \( d_{ij} \) is the distance between \( x_j \) and cluster center \( c_i \), \( U \) is a \( c \times n \) binary matrix called partition matrix. The individual elements

\[
u_{ij} \in \{0, 1\}
\]  

(1.2)

indicate the assignment of data to clusters: \( u_{ij} = 1 \) if the data point \( x_j \) is assigned to prototype \( C_i \), i.e., \( x_j \in \Gamma_i \); and \( u_{ij} = 0 \) otherwise. To ensure that each data point is assigned exactly to one cluster, it is required that:

\[
\sum_{i=1}^{c} u_{ij} = 1, \quad \forall j \in \{1, \ldots, n\}.
\]  

(1.3)

This constraint enforces exhaustive partitions and also serves the purpose to avoid the trivial solution when minimizing \( J_h \), which is that no data is assigned to any cluster: \( u_{ij} = 0, \forall i, j \). Together with \( u_{ij} \in \{0, 1\} \) it is possible that data are assigned to one or more clusters while there are some remaining clusters left empty. Since such a situation is undesirable, one usually requires that:

\[
\sum_{j=1}^{n} u_{ij} > 0, \quad \forall i \in \{1, \ldots, c\}.
\]  

(1.4)
depends on the two (disjoint) parameter sets, which are the cluster centers $c$ and the assignment of data points to clusters $U$. The problem of finding parameters that minimize the C-means objective function is NP-hard (Drineas et al., 2004). Therefore, the hard C-means clustering algorithm, also known as ISODATA algorithm (Ball and Hall, 1966; Krishnapuram and Keller, 1996), minimizes $J_h$ using an alternating optimization (AO) scheme.

Generally speaking, AO can be applied when a criterion function cannot be optimized directly, or when it is impractical. The parameters to optimize are split into two (or even more) groups. Then one group of parameters (e.g., the partition matrix) is optimized holding the other group(s) (e.g., the current cluster centers) fixed (and vice versa). This iterative updating scheme is then repeated. The main advantage of this method is that in each of the steps the optimum can be computed directly. By iterating the two (or more) steps the joint optimum is approached, although it cannot be guaranteed that the global optimum will be reached. The algorithm may get stuck in a local minimum of the applied objective function $J$.

However, alternating optimization is the commonly used parameter optimization method in clustering algorithms. Thus for each of the algorithms in this chapter we present the corresponding parameter update equations of their alternating optimization scheme.

In the case of the hard C-means the iterative optimization scheme works as follows: at first initial cluster centers are chosen. This can be done randomly, i.e., by picking $c$ random vectors that lie within the smallest (hyper-)box that encloses all data; or by initializing cluster centers with randomly chosen data points of the given data-set. Alternatively, more sophisticated initialization methods can be used as well, e.g., Latin hypercube sampling (McKay, Beckman and Conover, 1979). Then the parameters $C$ are held fixed and cluster assignments $U$ are determined that minimize the quantity of $J_h$. In this step each data point is assigned to its closest cluster center:

$$u_{ij} = \begin{cases} 1, & \text{if } i = \arg\min_{l=1}^{c} d_{lj} \\
0, & \text{otherwise} \end{cases}$$

Any other assignment of a data point than to its closest cluster would not minimize $J_h$ for fixed clusters. Then the data partition $U$ is held fixed and new cluster centers are computed as the mean of all data vectors assigned to them, since the mean minimizes the sum of the square distances in $J_h$. The calculation of the mean for each cluster (for which the algorithm got its name) is stated more formally:

$$c_i = \frac{\sum_{j=1}^{n} u_{ij} x_j}{\sum_{j=1}^{n} u_{ij}}.$$  

The two steps (1.5) and (1.6) are iterated until no change in $C$ or $U$ can be observed. Then the hard C-means terminates, yielding final cluster centers and data partition that are possibly locally optimal only.

Concluding the presentation of the hard C-means we want to mention its expressed tendency to become stuck in local minima, which makes it necessary to conduct several runs of the algorithm with different initializations (Duda and Hart, 1973). Then the best result out of many clusterings can be chosen based on the values of $J_h$.

We now turn to the fuzzy approaches, that relax the requirement $u_{ij} \in \{0, 1\}$ that is placed on the cluster assignments in classical clustering approaches. The extensions are based on the concepts of fuzzy sets such that we arrive at gradual memberships. We will discuss two major types of gradual cluster assignments and fuzzy data partitions altogether with their differentiated interpretations and standard algorithms, which are the (probabilistic) fuzzy C-means (FCM) in the next section and the possibilistic fuzzy C-means (PCM) in Section 1.2.3.

### 1.2.2 Fuzzy c-means

Fuzzy cluster analysis allows gradual memberships of data points to clusters measured as degrees in $[0,1]$. This gives the flexibility to express that data points can belong to more than one cluster. Furthermore, these membership degrees offer a much finer degree of detail of the data model. Aside from assigning a data point to clusters in shares, membership degrees can also express how ambiguously or definitely a data...
point should belong to a cluster. The concept of these membership degrees is substantiated by the definition and interpretation of fuzzy sets (Zadeh, 1965). Thus, fuzzy clustering allows fine grained solution spaces in the form of fuzzy partitions of the set of given examples \( X = \{x_1, \ldots, x_n\} \). Whereas the clusters \( \Gamma_i \) of data partitions have been classical subsets so far, they are represented by the fuzzy sets \( \mu_{\Gamma_i} \) of the data-set \( X \) in the following. Complying with fuzzy set theory, the cluster assignment \( u_{ij} \) is now the membership degree of a datum \( x_i \) to cluster \( \Gamma_j \), such that: \( u_{ij} = \mu_{\Gamma_j}(x_i) \in [0, 1] \). Since memberships to clusters are fuzzy, there is not a single label that is indicating to which cluster a data point belongs. Instead, fuzzy clustering methods associate a fuzzy label vector to each data point \( x_j \) that states its memberships to the \( c \) clusters:

\[
\mathbf{u}_j = (u_{1j}, \ldots, u_{cj})^T.
\]  

(1.7)

The \( c \times n \) matrix \( \mathbf{U} = (u_{ij}) = (\mathbf{u}_1, \ldots, \mathbf{u}_n) \) is then called a fuzzy partition matrix. Based on the fuzzy set notion we are now better suited to handle ambiguity of cluster assignments when clusters are badly delineated or overlapping.

So far, the general definition of fuzzy partition matrices leaves open how assignments of data to more than one cluster should be expressed in form of membership values. Furthermore, it is still unclear what degrees of belonging to clusters are allowed, i.e., the solution space (set of allowed fuzzy partitions) for fuzzy clustering algorithms is not yet specified. In the field of fuzzy clustering two types of fuzzy cluster partitions have evolved. They differ in the constraints they place on the membership degrees and how the membership values should be interpreted. In our discussion we begin with the most widely used type, the probabilistic partitions, since they have been proposed first. Notice, that in literature they are sometimes just called fuzzy partitions (dropping the word ‘probabilistic’). We use the subscript \( f \) for the probabilistic approaches and, in the next section, \( p \) for the possibilistic models. The latter constitute the second type of fuzzy partitions.

Let \( X = \{x_1, \ldots, x_n\} \) be the set of given examples and let \( c \) be the number of clusters \( (1 < c < n) \) represented by the fuzzy sets \( \mu_{\Gamma_i}, (i = 1, \ldots, c) \). Then we call \( \mathbf{U}_f = (u_{ij}) = (\mu_{\Gamma_i}(x_j)) \) a probabilistic cluster partition of \( X \) if

\[
\sum_{j=1}^n u_{ij} > 0, \quad \forall i \in \{1, \ldots, c\}, \quad \text{and}
\]

(1.8)

\[
\sum_{i=1}^c u_{ij} = 1, \quad \forall j \in \{1, \ldots, n\}
\]

(1.9)

hold. The \( u_{ij} \in [0, 1] \) are interpreted as the membership degree of datum \( x_i \) to cluster \( \Gamma_j \), relative to all other clusters.

Constraint (1.8) guarantees that no cluster is empty. This corresponds to the requirement in classical cluster analysis that no cluster, represented as (classical) subset of \( X \), is empty (see Equation (1.4)). Condition (1.9) ensures that the sum of the membership degrees for each datum equals 1. This means that each datum receives the same weight in comparison to all other data and, therefore, that all data are (equally) included into the cluster partition. This is related to the requirement in classical clustering that partitions are formed exhaustively (see Equation (1.3)). As a consequence of both constraints no cluster can contain the full membership of all data points. Furthermore, condition (1.9) corresponds to a normalization of the memberships per datum. Thus the membership degrees for a given datum formally resemble the probabilities of its being a member of the corresponding cluster.

Example: Figure 1.1 shows a (probabilistic) fuzzy classification of a two-dimensional symmetric data-set with two clusters. The grey scale indicates the strength of belonging to the clusters. The darker shading in the image indicates a high degree of membership for data points close to the cluster centers, while membership decreases for data points that lie further away from the clusters. The membership values of the data points are shown in Table 1.1. They form a probabilistic cluster partition according to the definition above. The following advantages over a conventional clustering representation can be noted: points in the center of a cluster can have a degree equal to 1, while points close to boundaries can be
identified as such, since their membership degree to the cluster they are closer to is considerably smaller than 1. Points on class boundaries may be classified as undetermined with a degree of indeterminacy proportional to their similarity to core points. The equidistant data point \( x_5 \) in the middle of the figure would have to be arbitrarily assigned with full weight to one of the clusters if only classical (‘crisp’) partitions were allowed. In this fuzzy partition, however, it can be associated with the equimembership vector \( (0.5, 0.5)^T \) to express the ambiguity of the assignment. Furthermore, crisp data partitions cannot express the difference between data points in the center and those that are rather at the boundary of a cluster. Both kinds of points would be fully assigned to the cluster they are most similar to. In a fuzzy cluster partition they are assigned degrees of belonging depending on their closeness to the centers.

After defining probabilistic partitions we can turn to developing an objective function for the fuzzy clustering task. Certainly, the closer a data point lies to the center of a cluster, the higher its degree of membership should be to this cluster. Following this rationale, one can say that the distances between the cluster centers and the data points (strongly) assigned to it should be minimal. Hence the problem to divide a given data-set into \( c \) clusters can (again) be stated as the task to minimize the squared distances of the data points to their cluster centers, since, of course, we want to maximize the degrees of membership. The probabilistic fuzzy objective function \( J_f \) is thus based on the least sum of squared distances just as \( J_h \).

<table>
<thead>
<tr>
<th>( j )</th>
<th>( x )</th>
<th>( y )</th>
<th>( u_{ij} )</th>
<th>( u_{ij} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>−3</td>
<td>0</td>
<td>0.93</td>
<td>0.07</td>
</tr>
<tr>
<td>1</td>
<td>−2</td>
<td>0</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>−1</td>
<td>0</td>
<td>0.94</td>
<td>0.06</td>
</tr>
<tr>
<td>3</td>
<td>−2</td>
<td>1</td>
<td>0.69</td>
<td>0.31</td>
</tr>
<tr>
<td>4</td>
<td>−2</td>
<td>−1</td>
<td>0.69</td>
<td>0.31</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td><strong>0.50</strong></td>
<td><strong>0.50</strong></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0.06</td>
<td>0.94</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0.01</td>
<td>0.99</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0.07</td>
<td>0.93</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>1</td>
<td>0.31</td>
<td>0.69</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>−1</td>
<td>0.31</td>
<td>0.69</td>
</tr>
</tbody>
</table>

**Table 1.1** A fuzzy partition of the symmetric data-set.